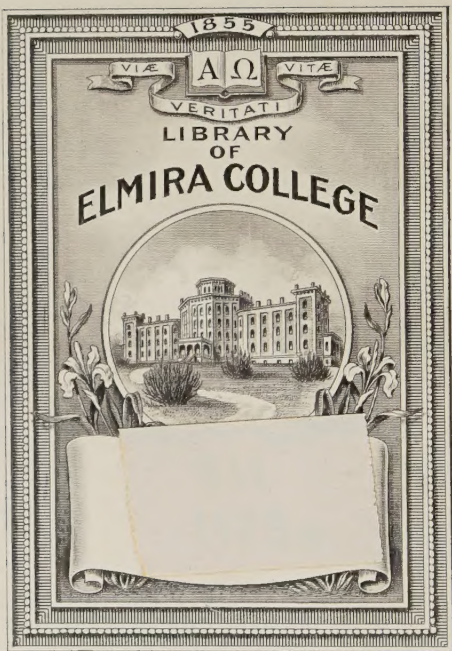












ELEMENTS OF PRACTICAL STATISTICS



THE MACMILLAN COMPANY  
NEW YORK • BOSTON • CHICAGO • DALLAS  
ATLANTA • SAN FRANCISCO

MACMILLAN & CO., LIMITED  
LONDON • BOMBAY • CALCUTTA  
MELBOURNE

THE MACMILLAN COMPANY  
OF CANADA, LIMITED  
TORONTO

H23

# ELEMENTS OF PRACTICAL STATISTICS

BY

1030 Henry  
F. H. HARPER, M.S., PH.D., M.AGR.

AGRICULTURAL ECONOMIST, BUREAU OF AGRICULTURAL ECONOMICS,  
UNITED STATES DEPARTMENT OF AGRICULTURE

FORMERLY ASSOCIATED WITH THE UNITED STATES TARIFF COMMISSION

ELMIRA COLLEGE LIBRARY  
ELMIRA, N. Y.

New York

THE MACMILLAN COMPANY

1930

COPYRIGHT, 1930,  
By THE MACMILLAN COMPANY

---

ALL RIGHTS RESERVED—NO PART OF THIS BOOK MAY BE REPRODUCED  
IN ANY FORM WITHOUT PERMISSION IN WRITING FROM THE PUBLISHER

---

SET UP AND ELECTROTYPED  
Published December, 1930

SET UP AND ELECTROTYPED BY T. MOREY & SON  
PRINTED IN THE UNITED STATES OF AMERICA.

HA  
29  
.H25



*To my wife, whose achievements have been  
an ever-increasing inspiration to me,  
this volume is affectionately  
dedicated*

36833

APR 7 1932



Digitized by the Internet Archive  
in 2023 with funding from  
Kahle/Austin Foundation

[https://archive.org/details/bwb\\_TW-360-499](https://archive.org/details/bwb_TW-360-499)

## PREFACE

This volume is the result of a long-cherished desire on the part of the author to present in concise and simple form some of the more common principles and methods involved in the technique of statistical analysis. This desire itself was the outgrowth of a failure to grasp the fundamental concepts underlying the true mathematics of elementary statistics when they were at first encountered. It is the feeling that, had the rudiments of analytical procedure been more fully understood when the earliest attempts were made to master the mathematics and theory of statistics, there would have been obviated a large part of the discouragement and tedious efforts that were to follow in later years. Because of these unfortunate experiences, and because it is felt that the student, or any other person, not trained in mathematics should be provided with a manual that contains in simple terms an explanation of the more important principles of statistics, the author has been prompted to prepare this book.

There are many treatises on statistical method, but most of them are either too advanced for the beginner or not sufficiently comprehensive within themselves to furnish the scope that is necessary for a clear and complete understanding of the fundamental phases of analytical technique. It is the purpose of this text to remove the mathematics of elementary statistics from the realms of theory and hieroglyphics and to pave the way for a ready grasp and keen appreciation of the more advanced phases of the subject. It is hoped that a perusal of the principles

and technique presented herein will serve to stimulate interest and further study. If only this is accomplished the writer will feel that his efforts have not been in vain.

In this volume the greatest emphasis has been placed upon analysis and the exact meaning and true significance of mathematical measurements. The opinion is that a great many students learn the mechanics of analytical methods without understanding the meaning and derivation of terms that are used, and by so doing they not only fail in the immediate purpose, but also in the ultimate aim. This has been constantly in the foreground throughout the preparation of this book, and attempts have been made to explain fully, yet briefly, the basis of all calculations and theoretical concepts involved. In doing this, special care has been given to logical sequence and the relationships of the various principles and analytical methods that have been considered. A most important phase in the study of statistics is that each factor or part be presented in its relation to the subject as a whole. It is only by understanding the intricate relationships that one can ever hope to become proficient in instituting analytical statistical studies and in drawing the proper conclusions from results obtained. This text purports to show the logical basis of these relationships in such a way that the entire field of statistics is seen to be composed of a series of closely associated parts. At the same time, while the text itself is not a complete compendium of statistical methods, the aim has been to include the more important essentials that are necessary for a comprehensive analysis of ordinary problems. It is believed that even without the aid of instruction the individual will find in this book a sufficient guide in making it possible to carry through to completion a great many of the analyses which it is desired to make.

Special attention is invited to one feature of this treatise. In the past it has been somewhat customary to present long narratives on method and significance in explaining the various phases of the subject. While there are many advantages in this, the author is of the opinion that too often the beginner is overwhelmed with long, drawn-out discussions, leaving only a hazy understanding of the matter under consideration. A better way, it would seem, is to state briefly and precisely the real meanings of terms and omit all discussion that is not necessary for ready comprehension. This text has been prepared on the basis of that principle, and the purpose in this connection has been to state in as few words as possible the true significance and fundamental concepts of the different statistical or mathematical measurements included within the scope of this work. The student, it is believed, will fully appreciate this feature as he progresses.

It would be a difficult task, indeed, to state fully the extent to which obligations are due to others. Like most present day workers, the author has profited greatly from many who have gone before, the results of whose tireless efforts have long been public property. In a few cases, however, the special character of the services rendered makes it quite necessary and proper that public expression be made of the obligation and indebtedness to those who have rendered these services. If there has been any success in completing the task which was undertaken it has been due in large measure to the assistance and helpful coöperation of Dr. William B. Kemp, Assistant Dean of the College of Agriculture, University of Maryland, whose interest in statistics has led him willingly to place a great amount of his time at the writer's service, and to contribute freely of his knowledge to the completion of this text. The in-

debtedness to Dr. Kemp assumes a dual aspect, since, in addition to assistance otherwise given, certain sections of this volume are borrowed in part from notes taken on his lectures in the classroom at the University of Maryland. It was while a student in Dr. Kemp's classes, during which time the subject of statistics was presented in such masterly manner, that the inspiration was first received to prepare a text of this sort, and because of this inspiration an exceptionally detailed compilation was made, with his assistance, of statistical terms and principles. These have been liberally drawn upon, either directly or as a source of suggestions. In addition to this an expression of indebtedness is due Dr. Kemp for information furnished at intervals from the time the actual preparation of this text was commenced. His kindly replies to inquiries have been a continuous and stimulating source of encouragement, without which this work would have been carried forward less rapidly, and much of the material contained herein would not have been presented at all if it had not been for his willingness to render assistance when the author's own ability was taxed beyond its capacity. It was because of his generous help that encouragement was given for renewed efforts from time to time, and which has made possible the completion of this text much earlier than at first anticipated. The author owes Dr. Kemp an everlasting debt of gratitude.

Acknowledgments are due also to Mr. A. J. T. Meurer, Cartographic Engineer, Economics Division, United States Tariff Commission. Under conditions which often necessitated sacrifices of time on his own part, Mr. Meurer labored untiringly in constructing the charts accompanying the text, and in many instances his expert judgment and criticism were utilized in perfecting graphic presenta-



tions, both of actual conditions and of illustrative material. From the very first consultation until the last chart was completed he diligently applied himself to the work, and from time to time improved the charts by valuable additions. Many sections of this text owe their completeness, or near completeness, to his assiduous care, and his suggestions at the proper time made it possible to pursue the work without interruption. Had it not been possible to secure such expert assistance as that rendered by Mr. Meurer the charts in this text could not have been presented in their proper form.

Acknowledgments are further made of the assistance received from Mr. William J. Kurtz, member of the agricultural staff of the United States Tariff Commission. He has read the more important chapters of the manuscript, and his numerous criticisms and suggestions have aided materially in eliminating certain discrepancies that had unintentionally been incorporated in it. The author feels particularly indebted to Mr. Kurtz for certain additions made to various sections of the text, all of which contributed to the completeness of the particular topic under consideration.

In some instances borrowings have been made from other texts. Many of these are specifically acknowledged by footnotes or by mention in the body of the text of which they constitute a part. Aside from these borrowings so mentioned, acknowledgments are due in other instances to Harry Jerome, Edmund E. Day, W. I. King, F. C. Mills, William A. McCall, and L. L. Thurstone.

The author is not unmindful of differences of opinion on certain phases of statistical technique. It may be that in using this text the reader will form conclusions as to methods of analysis and interpretation of results that are

not wholly in accord with those embodied herein. Perhaps, also, there may be certain parts of this treatise that some will think should have been discussed in either greater or less detail, or that more emphasis should have been placed upon particular principles. To those who direct attention to such factors, the author will feel extremely grateful.

F. H. HARPER

WASHINGTON, D. C.  
October, 1930

## INTRODUCTION

Statistics are becoming more and more important in the solution of most of our problems. In response to the demand for more definite knowledge of the changing conditions the federal, as well as the state and private agencies, have been constantly increasing their search for, and publication of, statistical facts. As a result it is becoming increasingly important for progressive business men, students, investigators, and legislators to understand the methods and the language of the statisticians. With the increased interest in statistics there have appeared several books dealing with this subject. Most of these books have, however, been intended principally for the use of statisticians, with the result that they are usually too complicated and involved for the ordinary reader. As already indicated, there is a large group of people whose work requires a general knowledge of statistical methods in order to interpret correctly the statistical material relating to the questions before them. This group is looking for a treatise upon statistics that explains in a simple way the general principles involved and one that is as free as practicable of technical terms and symbols.

It is the need for simplifying statistical terms that has prompted Dr. Harper to prepare this book, and it is my sincere hope that his efforts will help many to a better understanding of this important subject.

O. A. JUVE



# CONTENTS

|                                                | PAGE |
|------------------------------------------------|------|
| CHAPTER I. SAMPLING . . . . .                  | 1    |
| Uses . . . . .                                 | 3    |
| Method . . . . .                               | 8    |
| Representativeness . . . . .                   | 10   |
| Weighting . . . . .                            | 13   |
| Statistical Laws . . . . .                     | 24   |
| CHAPTER II. TABULATION . . . . .               | 31   |
| Classification of Data . . . . .               | 32   |
| Construction of Tables . . . . .               | 35   |
| CHAPTER III. FREQUENCY DISTRIBUTIONS . . . . . | 45   |
| Simple . . . . .                               | 51   |
| Cumulative . . . . .                           | 54   |
| Spatial . . . . .                              | 55   |
| Temporal . . . . .                             | 58   |
| Symmetrical . . . . .                          | 61   |
| Asymmetrical . . . . .                         | 63   |
| Normal . . . . .                               | 64   |
| CHAPTER IV. GRAPHIC PRESENTATION . . . . .     | 66   |
| Bar Charts . . . . .                           | 66   |
| Maps . . . . .                                 | 68   |
| Circle Charts . . . . .                        | 71   |
| Historigram . . . . .                          | 72   |
| Constituent-Element Chart . . . . .            | 74   |
| Silhouette Chart . . . . .                     | 75   |
| Excess and Deficit Chart . . . . .             | 77   |

| CHAPTER IV. GRAPHIC PRESENTATION ( <i>Cont'd</i> )   | PAGE    |
|------------------------------------------------------|---------|
| Ratio Chart . . . . .                                | 79      |
| Scatter Diagram . . . . .                            | 82      |
| Regression Chart . . . . .                           | 84      |
| Frequency Graphs . . . . .                           | 86      |
| Histogram . . . . .                                  | 86      |
| Frequency Polygon . . . . .                          | 87      |
| Smoothed Frequency Curve . . . . .                   | 88      |
| Diminishing Returns . . . . .                        | 89      |
| Highest Average Returns . . . . .                    | 91      |
| Increasing Returns . . . . .                         | 92      |
| Constant Returns . . . . .                           | 93      |
| <br>CHAPTER V. THE ARITHMETIC MEAN . . . . .         | <br>95  |
| Simple Mean . . . . .                                | 97      |
| Weighted Mean . . . . .                              | 99      |
| Mean of a Frequency Distribution . . . . .           | 106     |
| Progressive Mean . . . . .                           | 110     |
| Geometric and Harmonic Means . . . . .               | 110     |
| <br>CHAPTER VI. THE MODE . . . . .                   | <br>112 |
| <br>CHAPTER VII. THE MEDIAN . . . . .                | <br>117 |
| <br>CHAPTER VIII. VARIABILITY AND DISPERSION         | 122     |
| Frequency Table . . . . .                            | 123     |
| Frequency Graph . . . . .                            | 123     |
| The Range . . . . .                                  | 124     |
| Mean Deviation of a Series . . . . .                 | 124     |
| Mean Deviation of a Frequency Distribution . . . . . | 126     |
| Coefficient of Dispersion . . . . .                  | 128     |
| Standard Deviation . . . . .                         | 130     |



| CHAPTER VIII. VARIABILITY, ETC. ( <i>Cont'd</i> )            | PAGE    |
|--------------------------------------------------------------|---------|
| Standard Deviation from the Median . . . . .                 | 145     |
| Coefficient of Variation . . . . .                           | 145     |
| Quartile . . . . .                                           | 146     |
| Quartile Deviation . . . . .                                 | 150     |
| Deciles and Percentiles . . . . .                            | 151     |
| Skewness . . . . .                                           | 151     |
| Probable Error . . . . .                                     | 152     |
| <br>CHAPTER IX. PROBABILITY AND ERROR . . . . .              | <br>154 |
| Normal-Frequency Curve . . . . .                             | 154     |
| Application of the Probable Error . . . . .                  | 156     |
| Mean of a Sample . . . . .                                   | 156     |
| Single Variate . . . . .                                     | 159     |
| Standard Deviation . . . . .                                 | 159     |
| Coefficient of Correlation . . . . .                         | 160     |
| Distribution . . . . .                                       | 161     |
| <br>CHAPTER X. LONG-TIME TRENDS . . . . .                    | <br>162 |
| Free-Hand Trend . . . . .                                    | 163     |
| Moving Average . . . . .                                     | 164     |
| Semi-Average . . . . .                                       | 167     |
| Straight Line Trend of Least Squares . . . . .               | 168     |
| Logarithmic Parabola . . . . .                               | 178     |
| Second Degree Parabola . . . . .                             | 181     |
| Trends of Higher Order . . . . .                             | 185     |
| Price Deflation . . . . .                                    | 187     |
| <br>CHAPTER XI. LINEAR AND NONLINEAR RELATIONSHIPS . . . . . | <br>189 |
| Linear . . . . .                                             | 189     |
| Nonlinear . . . . .                                          | 193     |

|                                                                               | PAGE |
|-------------------------------------------------------------------------------|------|
| CHAPTER XII. CORRELATION . . . . .                                            | 195  |
| Simple . . . . .                                                              | 202  |
| Coefficient of Concurrent or Concomitant Deviations . . . . .                 | 203  |
| Pearsonian Coefficient . . . . .                                              | 205  |
| Coefficient of Correlation from Multiples of Standard Deviation . . . . .     | 208  |
| Coefficient of Correlation from Actual Values . . . . .                       | 211  |
| Correlation Based on Deviations from Straight Line Trend . . . . .            | 212  |
| Correlation of First Differences . . . . .                                    | 213  |
| Correlation by the Method of Percentage Change of First Differences . . . . . | 216  |
| Probable Error of Coefficient . . . . .                                       | 220  |
| Multiple . . . . .                                                            | 221  |
| Partial . . . . .                                                             | 228  |
| Rank-Order . . . . .                                                          | 231  |
| Correlation Index . . . . .                                                   | 233  |
| Correlation Ratio . . . . .                                                   | 234  |
| Correlation Table . . . . .                                                   | 237  |
| Frequency Distribution . . . . .                                              | 238  |
| Time Lag . . . . .                                                            | 246  |
| CHAPTER XIII. REGRESSION . . . . .                                            | 248  |
| Simple . . . . .                                                              | 250  |
| Multiple . . . . .                                                            | 256  |
| CHAPTER XIV. ESTIMATES AND ERROR . . . . .                                    | 262  |
| The Estimating Equation . . . . .                                             | 263  |
| Ratio Method of Estimating . . . . .                                          | 269  |
| Par Forecasting . . . . .                                                     | 272  |

# CONTENTS

xix

## CHAPTER XIV. ESTIMATES AND ERROR (*Cont'd*) PAGE

Standard Error of Estimate . . . . . 274

Coefficient of Alienation . . . . . 276

## CHAPTER XV. INDEX NUMBERS . . . . . 278

Simple . . . . . 279

Link . . . . . 280

Composite . . . . . 280

Accuracy . . . . . 283

## APPENDICES

A. Ordinates of the Probability Curve . . . . . 287

Areas of the Probability Surface . . . . . 290

B. Logarithms . . . . . 293

C. The Metric System . . . . . 299

D. Squares of Deviation from Points of Origin in

Straight Lines of Least Squares . . . . . 302

E. Measurements . . . . . 305

INDEX . . . . . 311



# ELEMENTS OF PRACTICAL STATISTICS



# ELEMENTS OF PRACTICAL STATISTICS

## CHAPTER I

### SAMPLING

In beginning the study of methods of statistical analysis the first problem to be considered is that of sampling. It is on the basis of material and data gathered by this process that many conclusions are drawn, and it is the only practical way by which an analytical statistical study can be made of the occurrences and conditions round about us that are too numerous or too widely scattered to be counted in their entirety. Sampling, then, is merely the taking of a part of a whole or total number of individuals, whatever they may be, from which to draw inferences or conclusions in regard to the characteristics of the group from which the sample was taken. A complete count, of course, is always preferable, but many times this is neither possible nor practicable, and it is largely because of these reasons that decisions and conclusions regarding the entire group must be based upon data selected by some phase or method of sampling.

There is a particular reason for discussing this important phase of statistical technique in the opening chapter of this text. An analysis cannot be made until the data are at hand from which to make it, and though this is so obvious it is no more obvious than the fact that results or



conclusions based on an improperly selected sample are often erroneous and misleading.

The first meeting of a class in biometry at a large university is vividly recalled. The instructor on that occasion was, and to-day is even more so, an exceptionally brilliant and resourceful scholar, with a rather remarkable ingenuity for presenting his subject. On this particular occasion, each member of the class was told to go to the near-by wood and make a collection of 200 leaves from the large number of young white-gum trees that were to be found there. No instructions were given as to which or how many trees the 200 leaves were to be taken from, it being assumed, no doubt, that at least one or two members of the class would, for one reason or another, bring back the desired sample representing the differences in color, size, and shape. When the class returned to the laboratory each member was told to separate the leaves according to color, placing the green and red ones in separate piles, the yellow ones in another, and the mixed colors in still another. When this was completed a count was made and the results recorded on the blackboard. To the interest of many, two samples were found to be made up almost entirely of yellow leaves, and another sample consisted of 160 green leaves and 40 leaves that were about equally divided among the other colors. Several samples brought in showed red to be the predominant color, but the great majority of them contained the several colors in almost constant proportions.

Upon questioning, the instructor brought forth the fact that those samples consisting of leaves that were nearly all of one color had been taken for the most part from only one or two trees, whereas the greater number of samples, consisting of leaves with the several colors in

fairly constant ratios, were found to have been taken from more than a score of the young trees. Then a count was made as to the size and shape, and there was the same evidence. The samples with the leaves that were mostly green or with leaves that were mostly yellow did not conform to the sizes and shapes of the leaves in the majority of samples.

This illustration of proper and improper sampling was readily appreciated, and its significance immediately grasped. No member of the class, perhaps, fully understood how to select a sample of leaves when the day was over, but a principle had been demonstrated which was not to be forgotten and which had served its purpose in showing the erroneous generalizations that might have been made on the basis of samples from a very limited number of trees. While this may not be a striking example of sampling, it has always been remembered because of the principle involved, and because of the demonstrative value.

### USES

There are many uses to be made of sampling in the fields of economics, education, sociology, and other social sciences. In fact, there is scarcely any limitation to be placed upon it. The investigator realizes even before he institutes an analytical study that scarcely ever is it practicable to take into consideration every individual part of the group or universe. A chemist, for example, is confronted with the task of determining the percentage of readily soluble nitrogen in a certain commercial fertilizer. To do this he does not attempt to make a quantitative analysis of all the fertilizer bearing that particular brand or label. All he does is to select a sample, a small quantity

of this fertilizer, analyze it, and on the basis of such facts as he ascertains draw his conclusions as to the entire lot. Were it not for the reliability of results obtained from this kind of sampling there would be a rather limited field for the practical application of analytical chemistry.

When the eastern farmer is deciding on the time to commence harvesting his wheat he does not make an examination or inspection of all the heads of grain, nor even of all parts of the field. He may not examine more than half a dozen heads, but that small number furnishes him the information he is seeking. If he finds that all the heads examined are in about the same stage of ripeness, and that sixty per cent., say, of the grains in those heads are in the dough stage, a conclusion is immediately drawn on the basis of that sample as to the stage of ripeness of the field as a whole, with a resulting decision as to when the harvesting is to commence.

The commercial cotton classer in the South does not examine all the lint in a bale when he is classifying the cotton in order more easily to assort it as to grade and staple length, or length of fibers. He merely takes small quantities of the lint, assuming that the lint is fairly representative of the bale from which it was taken. The sample may be quite small in relation to the size of the bale itself. Some classifications according to grade and staple length are made on the basis of samples of lint weighing only a few ounces, while in other instances the classers may prefer a sample weighing about a pound, or even more. The size of the sample may be made to vary in accordance with the apparent uniformity, or its lack, in grade and staple.

Further illustrations of the use of sampling may be presented by mentioning some practical applications.

1. *Farm cost data.* When it is desired to ascertain the average cost of producing short staple upland cotton on farms in North Carolina no attempt is made to study costs on every individual farm in the state. That would require such an expenditure of time and money as to make it altogether impracticable. All of such similar studies are based on samples, or sample groups of farms, which include cotton fields of all sizes, soils of all types and all degrees of fertility, and which represent other varying conditions. All the farm cost studies conducted by the United States Department of Agriculture are based upon this principle of sampling.

2. *Mileage tests for gasoline.* A large sales agency in Alabama made a mileage test with a well-known gasoline and a popular automobile. Measured quantities of gasoline were poured into a small specially constructed tank and the car driven over several different kinds of roads in the particular vicinity. The average mileage per unit of gasoline consumption was then used in an advertising campaign. It would not have been practicable, of course, to have driven over every road in the state, nor was it necessary. This is one of the cases in which sampling is highly satisfactory.

3. *High school grades.* Let us assume that an investigator in education wishes to know the average grade received by sixth-year grammar school pupils in Chicago schools for the month of January. Would he be likely to go to every sixth-grade teacher in the city, or to the central office and ask for such data? No, he would do nothing of the sort, unless he were exercising extraordinary caution and care. He would take the grades of pupils in a few schools in all sections of the city so as to be fairly certain that all nationalities, ages, abilities, and other

factors were properly represented. This would give a sample of all such grades in Chicago schools, and would afford the basis for drawing conclusions in regard thereto.

4. *Changes in food and clothing prices.* In attempting to show the changes in the general level of food and clothing prices no effort is made to record all the prices of all the articles that belong to this broad classification. Obviously this would require too much time, and, furthermore, the calculation of an average price would be difficult because prices in different places do not always apply to the same quantities, grade, and class of the particular article. Tomatoes in Texas may be quoted at 90 cents, which could possibly be for a crate, lug, or some other unit of measure, and at the same time on the Eastern Shore of Maryland they may be quoted at 40 cents, which would nearly always be the price for a basket of five-eighths bushel capacity, with a net weight ordinarily ranging from 30 to 35 pounds. Delaware cantaloupes are quoted by the United States Department of Agriculture at one dollar per crate as the average price for the 1928 season.<sup>1</sup> No one knows for a certainty what is meant by that price.

Those who have lived in the Delaware cantaloupe area know that there are crates of several different sizes used there. The flats, for example, are designed for 12 cantaloupes, whereas there are other crates made to contain 45, but these, themselves, are of several different sizes. There is the so-called standard crate, and then the larger sizes, which include crates of several dimensions. In that area a farmer never reports to his neighbor the price received without signifying in some way the size or kind of crate he has reference to, unless, of course, all this

<sup>1</sup> *Yearbook, United States Department of Agriculture, 1928, page 794, Table 177.*



is obvious. He may say that he received fifty, seventy-five, and a dollar, meaning fifty cents for the crates known as flats, seventy-five cents for the so-called standards, and one dollar for the larger crates, or jumbos, as they are popularly known. Even then the prices are not faultlessly designatory, for jumbo crates are made to contain both 36 and 45 cantaloupes, and, as implied above, these two crates are of different dimensions. In other words, there are several sizes of crates in which the same number of cantaloupes is packed, and other sizes in which different numbers are packed.

It becomes necessary, therefore, in determining the changes in general price levels to exercise careful judgment in choosing prices after it has been established as to how many units of each commodity or article are to be included. As a rule, those prices are selected which seem to be typical for the particular commodity, regardless of what they are or what units of measure they are based on. The student of price changes, then, must take a sample of all the many prices and from them calculate an estimate of the rise or fall of prices in general. Even in the selection of prices of specific commodities there must be a careful exercising of judgment. We may speak of the price of middling upland cotton on October 1, and yet not impart any information of value as to the cotton market for that day. Cotton may have sold for many different prices at many places on that day, and even at one place the price may have ranged from 10 to 15 cents. Here again the price that appears to be representative must be selected, and it is this selection that constitutes the sampling process.

5. *Juvenile delinquency.* The Children's Bureau, United States Department of Labor, in making a study of juve-

nile delinquency selected 40,000 cases in 42 courts in 15 states. An analysis of the records showed that "snitching apples" and "pranks of mischief" constituted the largest number of juvenile delinquency cases against boys. The conclusions thus drawn would probably have been no different even if a much larger sample had been drawn and made to include every juvenile court in the country.

### METHOD

There are no formulated methods of sampling that will apply equally to all kinds of data. The particular circumstances, such as location, purpose, ease of acquiring, length of time available, and size of sample desired will determine in large measure the method that will be devised or adopted in making the selection of individuals from a large number or group. Suggestions are offered as follows:

1. *Random or chance selection.* This method of selecting a sample is not based on any definite plan of choosing individuals from a group, but rather follows an aimless and undetermined course, with no apparent definite intention. If from a class of 100 pupils it is desired to choose 10 for some special purpose a good plan whereby each and every one of them will have an equal chance of being chosen is to write the names on separate pieces of paper of the same size, place the pieces of paper in a jar, and then thoroughly mix them. The 10 names may then be drawn at the same time from the top of the jar, or the names may be mixed after each individual is selected until the desired number has been drawn. In either case each name has an equal chance of being drawn, for the chances of one name being at the top of the jar are no greater and no less than the chances of all the others. This method of choosing a sample, therefore, is purely chance or random selection.

It may be applied in many cases, and there is much to be said in its favor. The implications so obviously involved are that the selection of individuals constituting the sample shall be based on principles of pure chance.

2. *Spatial or interval selection.* This method of selecting a sample can often be used to advantage. Let us assume that we are trying to find the average size of corn fields in Iowa, or the relation of corn acreage to the acreages of other crops. It would be a tremendous task to visit every farm and field in the state, so some way of sampling would necessarily be planned. A satisfactory method in many cases would be to select, say, every fifth, tenth, or twentieth farm along the roads in the areas covered by the inquiry. As a general rule it may be stated that the interval should be regulated by the distance traveled, or by the size of the area covered by the survey. If the study is to include the entire state perhaps even every thirtieth or fortieth farm or field would be sufficient for the purpose, but if only limited areas in selected parts of the state are to be canvassed then the numerical distance between farms and fields should be less.

This method could probably be applied also to the selection of samples in the making of soil-type studies. If, for instance, we wish to ascertain the predominant type or types of soil in a certain locality, and we have no knowledge of their composition and underlying strata, we may start from a fixed point and take samples at equal distances, the assumption being that such samples will be representative of the particular areas from which taken.

The spatial or interval method of sampling has a myriad of uses, as the preceding illustrations will indicate. If in a large city it is desired to estimate the average size of families a count may be made of every fifth house, say,



in each block, or of every twentieth house in each street. A school teacher may want a hasty approximation of the average grade of an incoming class. If the class is large, and if time for the moment is limited, an average may be calculated from the summation of every fourth or fifth grade as recorded in the register or roll book. A sample of telephone subscribers in Washington may be obtained by taking each hundredth name in the telephone directory, and a sample of taxpayers taken by selecting from a similar interval.

3. *Special selection.* Whenever the sampling process consists of choosing those individuals which appear to possess certain desired characteristics, and not necessarily the characteristics common to the entire group of which they are a part, the items or individuals may be said to be specially selected. An agricultural agent, for example, may wish to report on the results obtained or obtainable from certain applications of fertilizer to a specific crop. In choosing the individual farms on which to base his report or observations he may limit his selections to those that are better managed, and ignore those that do not meet certain requirements. Such a choice of individuals, then, would constitute a specially selected sample destined to serve a special purpose without indicating the general characteristics common, on the average, to the group as a whole.<sup>1</sup>

### REPRESENTATIVENESS

In any sampling process the selection of representative data must be the paramount aim. Samples that do not properly represent at least all of the important factors

<sup>1</sup> There are other methods of sampling, important among which is "selection by stratification." Since the general nature of stratification is indicated in the discussion of "weighting," it is purposely omitted here.

present in the universe <sup>1</sup> cannot be used with safety in drawing conclusions in regard to the group as a whole from which the sample was taken. Suppose an inquiry is to be made into the production costs of oranges in the United States. There are differences in cultural practices in the several states, differences in acreages under single management, differences in costs of bringing the trees to bearing age, and differences in yield. All these factors are directly related to unit cost of production. If we were attempting to determine, therefore, how much it costs on an average to produce oranges in the orange-growing states taken collectively the sample selected for such purpose should include acreages in all states where the industry is of any importance. In other words, the sample should represent the industry in its different phases, being so constituted as to include not only acreages in the several states, but also the varying factors that are present in each of them. The significance of this is too obvious to lend itself to further discussion.

As a similar illustrative principle, we may assume that the geneticist wishes to know the first generation inheritance of a specific characteristic in a plot of hybrid wheat. It will be immediately seen that all the heads or straws cannot be examined for this particular inheritance factor because of the great amount of work that would be involved. It becomes necessary, then, to select a sample from the plot. The geneticist in doing this will not purposely choose individuals that appear to him to be alike, but, on the contrary, will attempt to gather his material in

<sup>1</sup>The terms "universe" and "population" are used by statisticians as designations for the total number of items from which a sample is taken. If a sample of cotton gins were being taken to determine the percentage equipped with fire-fighting apparatus, the total number of cotton gins would be the universe.

such a way that all the differences will be represented. He will select both large and small heads, long and short stalks, and heads and stalks in varying stages of maturity. In addition to observing these factors he will give attention to the location of his individual selections. All the stalks and heads will not be taken from one corner of the plot because there may be some soil characteristic or some other influencing factor which has caused an unusual vegetative growth or stage of ripeness not readily discernible, either of which may be related to inheritance of the characteristic under study. To prevent erroneous and misleading conclusions that may be founded upon such sample, the selection of individuals is made from all parts of the plot so as to increase the degree of certainty that all influencing factors will be represented.

An attempt is sometimes made to draw a distinction between representativeness and adequacy. This is mere folly, however, and nothing of the sort is attempted in this text. The two may be used interchangeably, but the former is preferable to some because it is thought to be more suggestive. It is but an obvious fact that a representative sample is an adequate sample in the strict sense. The term "adequate" is often used in referring to the size of a sample, and in this way is distinguished from "representativeness." It is to be remembered by the student, however, that in order for a sample to be representative all the component attributes that are present in the group as a whole must be included in the smaller group of individuals selected to represent the whole. Furthermore, they must not only be present, but they must be present in the proper numbers and in the proper proportions. Thus the size of sample is one of the factors that constitute "representativeness."

An explanatory discussion on testing the adequacy and reliability of a sample will be found in a later chapter.<sup>1</sup>

### WEIGHTING

All samples, whenever it is possible and practicable, should be selected according to the ratio or proportion between important factors present, as has already been stated. Such a sample is a proportionately selected sample, with the items or factors in the sample properly weighted.

If it is desired to ascertain the average height of students in a coeducational high school or other institution the sample selected from which to estimate this average should be made up of boys and girls in the same proportion as the total numbers of each bear to one another. For example, let us assume that there are 2,500 students in the school, 1,500 of whom are boys and 1,000 of whom are girls. For every 3 boys there are 2 girls, or, in other words, the boys constitute 60 per cent. of the total number of students and the girls, 40 per cent. In taking a sample of these students from which to form an estimate of average height this same ratio should be maintained. It may be decided to calculate the average height of one-tenth of the entire student body of 2,500, giving a sample of 250 individuals. In keeping the same proportion or ratio in the sample between the number of boys and girls as there is in the group as a whole there would be selected 150 boys and 100 girls, 150 representing 60 per cent. of the sample and 100 representing 40 per cent. This would be a proportionate sample, with proper weight being given to the two factors under consideration, number of boys and number of girls. Other factors, of course, such as age, grade in school, etc., may be taken into considera-

<sup>1</sup> See the discussion on "Probable Error of the Mean," pages 156 to 159.

tion also in the sampling process. This involves a little more complication, however. Suppose the 2,500 students are distributed among the four classes as shown in the following table.

TABLE I. DISTRIBUTION OF HIGH SCHOOL STUDENTS BY CLASSES <sup>1</sup>

| CLASS     | NUMBER OF STUDENTS<br>IN EACH CLASS | PER CENT. EACH CLASS IS OF<br>THE WHOLE NUMBER |
|-----------|-------------------------------------|------------------------------------------------|
| Freshman  | 875                                 | 35                                             |
| Sophomore | 750                                 | 30                                             |
| Junior    | 500                                 | 20                                             |
| Senior    | 375                                 | 15                                             |
| Total     | 2,500                               | 100                                            |

We already know that of the 2,500 students, 1,500 are boys and that 1,000 are girls, and we have seen how to select a proportionate sample with only those two factors considered or weighted. It is desired to improve or increase the degree of accuracy by taking into consideration the differences in relative proportions of boys and girls in the individual classes. Table I shows the number of students in each of the four classes, and also the per cent. that the number in each class is of the total number in the school. There is, as we have seen, a ratio of 3 boys to 2 girls, or 60 per cent. boys to 40 per cent. girls for the school as whole, but there is not this same ratio between the number of boys and girls in the individual classes. The table on page 15 illustrates this.

We see in the following table that although for the four classes as a whole the ratio between the number of boys and girls is 60 to 40, measured in percentage, or 3 to 2, as measured in actual numbers, there is a different ratio between the number of boys and girls in individual classes.

<sup>1</sup> Hypothetical data.



TABLE II. DISTRIBUTION OF HIGH SCHOOL BOYS AND GIRLS BY CLASSES <sup>1</sup>

| 1         | 2               | 3                                                           | 4               | 5                                                           | 6                            |
|-----------|-----------------|-------------------------------------------------------------|-----------------|-------------------------------------------------------------|------------------------------|
| CLASS     | Boys            |                                                             | Girls           |                                                             | RATIO OF<br>BOYS<br>TO GIRLS |
|           | Total<br>Number | Per Cent. of<br>Total of All<br>Boys and Girls<br>in School | Total<br>Number | Per Cent. of<br>Total of All<br>Boys and Girls<br>in School |                              |
| Freshman  | 475             | 19                                                          | 400             | 16                                                          | 54:46                        |
| Sophomore | 450             | 18                                                          | 300             | 12                                                          | 60:40                        |
| Junior    | 325             | 13                                                          | 175             | 7                                                           | 65:35                        |
| Senior    | 250             | 10                                                          | 125             | 5                                                           | 67:33                        |
| Total     | 1,500           | 60                                                          | 1,000           | 40                                                          | 60:40                        |

In the freshman class there are 475 boys and 400 girls, or a ratio of 54 boys to 46 girls, which is the same as stating that 54 per cent. of the total of 875 students in the freshman class are boys and 46 per cent. are girls. Of the sophomores, there are 60 boys to 40 girls; of the juniors, 65 boys to 35 girls; and of the seniors, 67 boys to 33 girls. The problem now is to select a sample containing 10 per cent. of the total number of all students in the four classes, and at the same time to weight the items or factors in the sample both as to classes and as to the ratio or proportion between the number of boys and girls in each class. This will be seen to be a very simple process.

Let us refer again to Table I, recalling that 10 per cent. of the total number of pupils, or a sample of 250, is to be selected. It will be seen that of this 250, 35 per cent. is to be taken from the freshman class, 30 per cent. from the sophomore class, 20 per cent. from the junior class, and 15 per cent. from the senior class, which are the respective totals of the figures given for each class in the third and fifth columns in Table II. This means that of the 250 students selected, 87 will be freshmen, 75 will be sopho-

<sup>1</sup> Based on data in Table I, page 14.

mores, 50 will be juniors, and 38 will be seniors, which is the same as taking 10 per cent. of the number in each class as shown in Table II. Next it must be determined how many girls and how many boys are to be selected to make up the total of 87 freshmen, 75 sophomores, 50 juniors, and 38 seniors. Turning to Table II and reading under the heading "Ratio of boys to girls" in column 6 we find that of the freshmen selected, 54 per cent. are to be boys and 46 per cent. are to be girls. Of the sophomores, 60 per cent. are to be boys and 40 per cent. are to be girls, and so on for the juniors and seniors. The following table will make the calculations more easily understood.

TABLE III. SAMPLING OF HIGH SCHOOL STUDENTS BY CLASS AND SEX <sup>1</sup>

| CLASS     | TOTAL NUMBER<br>OF STUDENTS TO<br>BE SELECTED<br>FROM EACH<br>CLASS | BOYS TO BE SELECTED          |                           | GIRLS TO BE SELECTED         |                           |
|-----------|---------------------------------------------------------------------|------------------------------|---------------------------|------------------------------|---------------------------|
|           |                                                                     | Per Cent. from<br>Each Class | Number from<br>Each Class | Per Cent. from<br>Each Class | Number from<br>Each Class |
| Freshman  | 87                                                                  | 54                           | 47                        | 46                           | 40                        |
| Sophomore | 75                                                                  | 60                           | 45                        | 40                           | 30                        |
| Junior    | 50                                                                  | 65                           | 33                        | 35                           | 17                        |
| Senior    | 38                                                                  | 67                           | 25                        | 33                           | 13                        |
| Total     | 250                                                                 | —                            | 150                       | —                            | 100                       |

The calculation as now made shows us that in order to improve or increase the degree of accuracy in drawing a sample of high school students to serve as the basis for estimating the average height of the entire group we must not only select the desired number of individuals, 250 in this case, but there must be some consideration given to the ratio or proportion between the number of boys and girls in the school, and also to the ratio between the number of boys and girls in the individual classes. As calculated in Table III the factors in the sample are weighted for

<sup>1</sup> Based on data in Tables I and II, pages 14 and 15.



the differences in ratios between the number of boys and girls in each class. If we were selecting a proportionate sample and weighting only for the ratio between the number of boys and girls in the school as a whole all that would be necessary, recalling that 60 per cent. of the student body is made up of boys and 40 per cent. of girls and that the sample is to consist of 250 individuals, would be to take 60 per cent. of 250, which is 150, and 40 per cent. of 250, which is 100. These figures represent the numbers of boys and girls to be selected, but there is nothing to indicate whether they are to be taken from the freshman class, from a single one of the other classes, or whether they are to be taken in equal numbers from each of the four classes. Table I shows what per cent. is to be selected from each class, and Table III shows the actual numbers of boys and girls needed to make up the desired sample. This is one of the instances in which the items in a proportionate sample can be weighted fairly accurately for several factors.

Another method may be followed in arriving at the results in Table III, and the student of statistics will probably adopt it as he progresses. To demonstrate the principle, however, and to show the details in the process of selecting a proportionate sample and of weighting the items in the sample when two or more factors must be considered, it is desirable that a thorough mastery be gained of the simple mathematics involved in Tables I, II, and III.

The weighting of items or factors in a proportionate sample is particularly desirable when there are marked differences in the individuals constituting the group from which the sample is to be taken. Let us illustrate the principle further. In Baltimore there are many thousands

of automobiles, ranging in original cost or value from a few hundred dollars each to several thousand dollars each. If we were to go to the records of the commissioner of motor vehicles and select a sample of these automobiles on which to base an estimate of the average of original cost of all the automobiles in the city, the degree of approach to accuracy of our estimate would be increased by taking into consideration the relative proportion that each make of car of different value constitutes of the total number registered. Let us assume there are 100,000 cars titled in the city, and that they fall in the following classifications: 20,000 originally costing \$700 each; 30,000, \$1,200 each; 15,000, \$1,800 each; 10,000, \$2,500 each; 18,000, \$3,000 each; 5,000, \$4,000 each; and 2,000 costing \$5,000 each. A sample of these automobiles chosen without regard to the ratio that the number of cars of different values bears to the total number of 100,000 would not be likely to give the desired estimate of average value or original cost. Obviously, then, the sample selected in such a case is subject to criticism when the factors characterizing the universe are unweighted simply because no account has been taken of the fact that the number of automobiles of the several values bear different ratios to the total number. The 20,000 automobiles with an original cost of \$700 each constitute 20 per cent. of the total 100,000, the 30,000 with an original cost of \$1,200 each constitute 30 per cent. of the total, and so on for the other groups of 15,000, 10,000, 18,000, 5,000, and 2,000. Now, if it is desired to select a sample of one-fifth of the 100,000 we would have in it a total of 20,000 automobiles, but this alone does not indicate whether the 20,000 should be selected from one group, or whether they should be selected in equal number from the seven groups. The ratio

between the number of automobiles in each group and the total number of 100,000 decides this for us. The 20,000 automobiles of the \$700 group constitute, as has been stated, 20 per cent. of the total, and the number selected from this group will, therefore, constitute 20 per cent. of the sample. The automobiles of the \$1,200 group will furnish 30 per cent. of the sample selected, and so on for the other groups. This method of choosing individuals will give a proportionately selected sample with the varying factor of original cost or value properly weighted, and it is the only way of selecting a sample that will give the desired estimate of average cost or value.

Weighting of items or factors in proportionate samples is sometimes attempted in the making of farm cost studies. It is decided, for instance, that an inquiry is to be made into farm costs of producing winter wheat in the area comprising the states of Kansas, Nebraska, Missouri, and Illinois. The first problem that must be met is the making of a decision as to how many farmers are to be interviewed in regard to costs of production and what part of the total acreage that costs are to be obtained for. After some study as to estimated acreage and number of farmers growing wheat in those states and the average acreage of wheat per farm, it is decided, we will say, to interview 300 growers, representing 5 per cent. of the acreage harvested, or, in other words, to obtain 300 separate records of costs. The next decision to be made is in regard to the number of growers to be interviewed in each state and the acreage in each state for which costs are to be obtained. This is quite a simple matter to decide, for we have already seen how a proportionate sample, with proper weight being given to the characterizing factors, is selected in such instances. The total acreage

of winter wheat harvested in 1928 in the four states named was 16,682,000 acres.<sup>1</sup> Of this acreage, that of Kansas represented approximately 62 per cent., as compared with 21 per cent. for Nebraska, 9 per cent. for Missouri, and 8 per cent. for Illinois. For sake of simplicity we will assume that the number of farmers producing wheat in the several states bears the same relation to the total number as the acreage harvested in each state bears to the total acreage harvested in all the states under consideration. The sample as selected, therefore, should be constituted in those proportions. That is to say, 62 per cent. of the 300 records, as regards both number of farmers and size of acreage, should be taken in Kansas, 21 per cent. in Nebraska, 9 per cent. in Missouri, and 8 per cent. in Illinois. This will give a proportionate sample for 1928 of winter wheat acreage in the area comprising the states named, with the factors of acreage and number of farmers properly weighted. If costs are to be taken for two or more years the sample may be based on the ratio between the average acreage in each state and the average of total acreage in all four states for the same period, with the same consideration given to the number of growers as though only one year were considered.

Another method of selecting the sample of winter wheat growers would be on the basis of production in each state as compared with the total production in the four states. This is not as satisfactory, however, as selecting the sample on the basis of acreage, though it is frequently used. The big disadvantage in selecting a sample of such farm data on the basis of production rather than on the basis of acreage is due to the differences in yield. A sample selected for purposes of a cost

<sup>1</sup> *Yearbook, United States Department of Agriculture, 1928*, page 672, Table 3.

inquiry should, if it is to be a proportionate sample, be based on the acreage, therefore, in order that the average costs may show the highest degree of approach to accuracy. Cost inquiries based on samples selected only in accordance with ratios between production in the individual states and total production for all the states to be surveyed are likely to show biased results because of the usual close relationship of yields to costs.

Aside from weighting the items in a sample of farms and acreage selected for a cost inquiry on the basis as described above, there are other factors that may be considered. Important among these are the differences in acreage on individual farms or under single management, types of soil, labor conditions, distance from market, and systems of tenure. In an area where 50 records are to be taken, let us say, it may be that on some of the farms the acreages of the crop being studied are small in comparison to the acreages of other crops, whereas on other farms the production of this particular crop may be the major enterprise. Obviously, this factor, if practicable, should be considered in selecting a sample. The same may be said of differences in distance from market and of differences in other factors. It will be seen that the proper weighting of items in a sample in a farm cost inquiry is rather difficult, and, at the best, weighting can be done for only a few of the most important of them.

Scientific investigators are sometimes inclined to overestimate their process and details of sampling when they list the number of varying factors taken into consideration in selecting their samples for purposes of study. The long list of factors enumerated may have been taken into consideration in one way or another, but it is quite a safe assertion to make in many cases that where several of



them are present in the universe in different proportions the items in a sample, particularly the items in a large sample made up of several smaller ones, as selected are properly weighted for not more than one or two of the most important of them, regardless of what may be said otherwise in regard thereto. One of the principal reasons for this, of course, is the frequent impossibility of selecting a sample in which items or characteristic factors are truly weighted, many of which may not be readily discernible in their actual and relative importance.

The preceding discussion of sampling under conditions where several factors are concerned is designed largely to imply that weighting, important as it is, must not be carried to an absurdity. Each and every selection of a sample has its own problems, and these must be decided by the investigator himself. A thorough mastery of the principles involved and a trained judgment as to what factors should be considered are the essentials for the sampling of any mass of data. In the final analysis the paramount purpose of proportionate sampling and proper weighting of items in a sample is to obtain a true average of conditions, or an average that is as nearly true as possible under prevailing circumstances. Further reference to this will be made in this chapter and also in Chapter V, where it is shown that in reality a weighted average is merely a simple average correctly calculated.

In the strict interpretation of terms, a truly representative sample is a proportionately selected sample, with all the items or factors properly weighted, for unless all the characteristic features of the universe are included the sample does not properly represent the group or universe from which it was taken. Generally, the two terms "representativeness" and "weighted" are most likely to be



confused when a big sample is made up of two or more smaller samples. In this case each of the individual samples may be quite representative of the groups from which selected, and still the composite sample made up of the several selections may not be proportionate. Perhaps an illustration will make this more easily understood. Suppose we are determining from sample data the average age of the combined populace of Baltimore and Washington. A representative sample of Baltimore may be made up of 25,000 people, including inhabitants of all ages in exactly the proper proportions. The sample of Washington residents may also consist of 25,000 people of varying ages directly proportionate to the total population of the city. This sample, too, would be representative, but there is a difference in the total population of the two cities, so that the 50,000 individual ages cannot be considered as representative of the combined population. In other words, the sample of 50,000 would not be proportionate because the ratio of the size of the two individual samples to the size of the groups from which selected is not the same in each case. A simple arithmetic mean of the 50,000 ages would not be expected to represent the true average age of the combined population. In order for this to be true it would be necessary not only for each of the samples to be truly representative of the larger group from which it was taken, but also that the relative magnitudes of the two samples be the same as the relative magnitudes of the total populations in the two cities. It is in that instance that the simple average would be identical with the weighted average. Thus it is that the weighted average is merely the simple average correctly calculated. Sometimes it may be easier and more convenient to weight the final results, ignoring the factor of

proportionate sampling in respect to the component parts of a large composite sample. For example, if the average age of Baltimore residents is 44 years, and the average age of Washingtonians is 39 years, a weighted average may be obtained by giving to each of these averages a weight in proportion to the relative numbers of people living in the two cities. The 1927 estimates of population show 819,000 for Baltimore and 540,000 for Washington.<sup>1</sup> On the basis of these figures the population of Baltimore in 1927 was 1.5 as great as that of the City of Washington. To calculate the weighted average age of the two populations it is but necessary to apply the weight of 1.5 to the average age for the Baltimore population and the weight of 1 to the average age of Washingtonians, summate the products obtained by multiplication, and then divide by the sum of the weights. Thus we have 44 multiplied by 1.5, giving a product of 66, and 39 multiplied by 1. The sum of 66 and 39 is 105, which divided by the sum of the weights (1.5 plus 1) gives 42, the weighted average age.

### STATISTICAL LAWS

From the preceding discussion the student will see that all sampling is an attribute of the regularity with which certain factors occur, and that it is based on the assumption that a part of a total group properly selected tends to have, and usually does have, on the average, the same essential characteristics as the group from which the sample was taken. If we stand for an hour, for example, at Seventh Street and Pennsylvania Avenue in Washington and note the color of hat worn by every fifth man who passes by the chances are we will have a fairly true indi-

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 43, Table 39.

cation of the color of hats worn by all the men who passed during that time, and by modifying the count an indication of the color of hats worn by men in Washington as a whole could be obtained. In like manner a count made of the crop acreage in every tenth farm in Maryland's truck-growing area will give an average acreage per farm that is likely to be very close to the average for all the farms.

Perhaps a better illustration than either of the foregoing may be given of farm costs. The assumption underlying all cost inquiries is that the average cost as calculated on the basis of samples selected is the same, or nearly the same, as the average cost for the whole group of farms from which the sample was taken. Thus it is that in making such inquiries we must necessarily assume that the average cost for the sample selected is about the same as the average for any other like sample. It follows, then, that a properly selected sample usually gives, on the average, the same results that would be obtained if the whole group were to be considered from which the sample was taken.

The general rule as described does not necessarily imply that the characteristics of a sample selected from a large group will, in every case of single samples, be exactly the same as those of the group as a whole. There may be some variation due to chance regardless of how much care and precision are exercised in sampling, but these variations would tend to disappear if a sufficiently large number of samples were taken and the average calculated. This phase of regularity involves the principle of probability and error, discussion of which must be delayed for a later chapter.<sup>1</sup>

<sup>1</sup> Chapter IX, pages 154 to 161.

In a single sample there is sometimes found an individual with unusual characteristics, or one with characteristics that are greatly different from those of corresponding factors associated with other individuals constituting the sample. An example of this may be given in farm cost inquiries. Among a group of 25 growers selected for purposes of interviewing in regard to costs of producing a specific commodity it may be found that on one farm a single item of costs, feed for work stock, perhaps, is three times as great per animal as the same cost on the farm ranking second in order in respect to this particular cost. Other samples selected in the same way as the first would be expected to show this same single characteristic in the same proportion, since if the first sample shows the characteristics of the group from which it is taken, then another sample similarly selected should be essentially the same as the first one. Statisticians have designated this tendency or principle as "the permanence of small numbers," meaning that when a particular unusual characteristic occurs in a properly selected sample it may be expected that this same characteristic is likely to be present in the entire group from which the sample was taken in the same proportion as it is present in the sample itself. In other words, the characteristic is permanent, being present in every group similarly selected and composed of the same number of individuals as the original sample.

The fundamental principle underlying the selection of a sample when there are great variations in the same factors that characterize the group from which the sample was taken is the tendency for differences to equalize each other. That is, the extent to which an item is overestimated is likely to be offset by a corresponding under-

estimate. An illustration will make this clear. One hundred New York farmers are to be interviewed as to how much it costs them to produce fluid milk. We will assume that, as is usually the case, complete cost records are not maintained. For this reason, therefore, the farmer's estimate of certain items of costs must sometimes be based on his memory or judgment extending over a period of a year or more and on his estimate as to the importance of the particular item. Suppose this item of cost is water for drinking purposes. There may be a pumping outfit that furnishes water for domestic purposes in the home, for dairy cattle, for stock other than dairy cattle, and for watering lawns. The farmer could probably give a fairly accurate estimate of the total annual cost of the pumping outfit, including repairs, depreciation, fuel, etc., but his allocation of this total cost to the different purposes for which the water was used would be more difficult. There may be some who would overestimate the proportion of total cost that should be charged to dairy cattle, whereas others would likely underestimate the cost. The result would be that one error would tend to offset the other, so that in the final analysis the average cost of providing water as ascertained from the records of the farmers interviewed would likely be the same or nearly the same as the average cost furnished by other groups similarly selected and interviewed, and thus the average would be considered as being representative for the entire group from which the 100 farmers were selected.

This tendency for numbers to offset one another is known by statisticians as the "inertia of large numbers," and were it not for this tendency the sampling of a universe in many cases would not furnish a satisfactory basis for drawing conclusions in respect to a large group from the



characteristics of a smaller group selected from it. The student will bear in mind that when there are involved any estimates that are based on memory or judgment alone one of the essential things for the investigator to do is to select a sample sufficiently large for the overestimates and underestimates to offset one another.

Before passing on to the next statistical law a modification must be made of what has been stated in respect to the inertia of large numbers. Whenever there is present any condition or influence that causes, or even tends to cause, a constant increase or decrease in any particular factor the offsetting tendencies of large numbers cannot always be relied upon to remove the effects of individual inaccuracies. The levying of an increased duty on agricultural products, for example, might conceivably cause an increase in unit cost of production in the country levying the duty, if production were stimulated thereby, and assuming producers of this commodity to be already producing at the lowest average cost, with the consequence that any increased production must be accompanied by an outlay in additional expenditure that is greater in proportion to the increased production than the relationship formerly existing between expenditure and production. In this case a duty that increased as production increased would tend to be associated with a constantly increasing unit cost of production of the particular commodity. This is one phase of costs that is often likely to be overlooked by high-tariff advocates, particularly by high-tariff advocates in agricultural communities, though it is a pronounced statistical and economic law. A duty, for example, that equalizes the low production cost in a foreign country and the high domestic cost might eventually be advantageous to the foreign competitor because



his production may not yet have reached the point of highest average returns, whereas any increased production on the part of the domestic producer may be brought about by a unit cost greater than the unit cost prior to the levying of the increased duty.

One of the most important of mathematical laws having application to statistical technique is the one pertaining to the relationship between the size of sample and its degree of precision, or the extent to which it is truly representative. It is obvious, of course, that the reliability of a sample, or the extent to which a sample may be expected to reflect the characteristics of the group from which it was taken, depends somewhat upon the size of the sample. Mathematicians have demonstrated that the degree of precision, increasing as the number of items in the sample is increased, varies in accordance with the square root of the number of items constituting the sample. If we have taken a sample of 25 farm cost records, let us say, to serve as a basis for estimating the average cost of producing a particular commodity and wish to increase the degree of precision by 100 per cent. it will be necessary to increase the sample to four times its original size. The square root of 25, we see, is 5. Now, to double that number, or to increase it by 100 per cent., a product of 10 is obtained, which is the square root of 100. It would require, therefore, a sample of 100 records to double the degree of precision of a sample of 25. Similarly, if the degree of precision of the sample is to be trebled the sample would need to be nine times as great as the original one selected. The mathematics involved here is simple. We have just seen that 5 is the square root of 25, and that 10 is the square root of 100. Increasing 5 to three times its size we obtain a product of 15, which is the square

root of 225, meaning that 225 records, or a sample nine times as great as the original sample of 25, would need to be obtained in order to increase the degree of precision by 3. A sample of 400 individuals would be necessary to quadruple the degree of precision, and 625 to increase it by 5. The student will grasp the significance of this when he takes up the mathematics underlying the probable error and the relation of the magnitude of the probable error to the number of items involved.

## CHAPTER II

### TABULATION

When a mass of data has been assembled the first step to be taken in the process of analyzing and interpreting it consists of some sort of concise or compact arrangement. You have made a survey, we will assume, showing acreage planted to each specific crop on 50 farms in Georgia, the acreage of crop land devoted to pasture, and the acreage of idle land. The data collected show the number of acres of crop land in each of the farms covered by the inquiry, and in addition there is shown the acreage of the different crops on each specific farm. In other words, you have 50 complete individual records of the acreages of crops grown on 50 farms. Before this information can be of much value it is necessary that like or similar component parts be grouped under single headings. It is desired to know, for instance, how many acres of cotton were grown on the 50 farms, or how many acres of corn, or how much of the crop land was devoted to temporary pasture. These can best be ascertained by recording these acreages for each farm under properly designated headings and then adding, or summing, the individual numbers to obtain the total for the entire group of farms. This procedure is known by statisticians as tabulation, which is merely the arrangement of data into tables, or in tabular form, the most elementary phase of statistical technique.

## CLASSIFICATION OF DATA

Before data of any kind are tabulated there must be some decision made as to the classes into which the arrangement or tabulation will be made. The 50 records referred to above show for individual farms the acreages of the different crops, the acreage of pasture, and the acreage of land that was idle during the year. In making up a table in which to arrange these different acreages in tabular form some classification must be made in order that each figure may be recorded in its proper place, or under the appropriate heading. It may be that the crop land will be classified into three groups, or under three headings, acreage planted to crops, acreage in pasture, and acreage remaining idle. There would then be three headings under which the total acreage would be tabulated. In other words, there would be three columns of figures. Each farm would be taken separately and the acreage designated by the classification recorded under the proper heading.

Perhaps a more detailed tabulation would be wanted. It may be, as stated earlier, that the acreage of each individual crop is to be recorded separately. In this case the acreage planted to crops would be divided into subclasses, such as acreage in cotton, acreage in tobacco, etc., and the figures representing these acreages recorded under these headings.

The identical procedure would be followed in tabulating any other similar mass of data. The president of a large corporation with several branches or subsidiaries may wish to know the exact number of different classes of employees in the plants under his management. His assistants in preparing this information would construct a table, and under the appropriate headings would record

for each plant the number of employees belonging to each classification. The following table, based on assumed data, is presented for illustrative purposes.

TABLE IV. CLASSIFICATION OF EMPLOYEES IN RICE MILLS <sup>1</sup>

| LOCATION OF<br>MILL | CLASSIFICATION AND NUMBER OF EMPLOYEES |                |                      |                 |                     | TOTAL |
|---------------------|----------------------------------------|----------------|----------------------|-----------------|---------------------|-------|
|                     | Man-<br>agers                          | Over-<br>seers | Salaried<br>Laborers | Wage<br>Earners | Contract<br>Workers |       |
| New Orleans         | 1                                      | 4              | 14                   | 25              | 10                  | 54    |
| Houston             | 1                                      | 3              | 12                   | 23              | 6                   | 45    |
| St. Louis           | 1                                      | 3              | 8                    | 19              | 2                   | 33    |
| Lake Charles        | 1                                      | 2              | 7                    | 12              | 1                   | 23    |
| Total               | 4                                      | 12             | 41                   | 79              | 19                  | 155   |

The data presented in this form permit one to see at a glance the number of employees of each class that are engaged at each mill, and the arrangement is such as to afford a ready comparison of the numbers in each group. This is *qualitative* classification. The wage earners, for example, are not necessarily employees of a rice mill merely because they are wage earners, but they are wage earners because they are employees of the mill. There is no other place for them in the establishment other than the position of wage earner. In another place they might be contract workers, or even overseers.

Data may be classified according to geographical location. The agricultural economist is interested in knowing not only the total acreage of wheat in the country, but he wishes to know also the acreage in each state. In other words, he wants to know where the acreage of wheat is located, or in what parts of the country it is grown. This may be designated as *geographical* classification. Similarly, it may be desired to know the location of manufacturing industries, as illustrated in the following table.

<sup>1</sup> Assumed data.

TABLE V. SUMMARY OF ALL MANUFACTURING INDUSTRIES IN THE UNITED STATES FOR 1925, BY GEOGRAPHICAL DIVISIONS <sup>1</sup>

| GEOGRAPHICAL DIVISION | NUMBER OF MANUFACTURING ESTABLISHMENTS |
|-----------------------|----------------------------------------|
| New England           | 18,173                                 |
| Middle Atlantic       | 58,895                                 |
| East North Central    | 42,888                                 |
| West North Central    | 16,280                                 |
| South Atlantic        | 16,576                                 |
| East South Central    | 8,080                                  |
| West South Central    | 7,887                                  |
| Mountain              | 3,849                                  |
| Pacific               | 14,762                                 |
| Total                 | 187,390                                |

The preceding table shows not only the number of manufacturing establishments in the United States, but also the geographical sections of the country in which these establishments are located. A further classification could be made as to the location by states, and even by counties within the state. This would still be *geographical* classification.

TABLE VI. DISTRIBUTION OF TOTAL POPULATION IN THE UNITED STATES BY AGE GROUPS FOR THREE DECADES <sup>2</sup>

| AGE GROUP<br>(Years) | NUMBER     |            |             |
|----------------------|------------|------------|-------------|
|                      | 1900       | 1910       | 1920        |
| Under 5              | 9,170,628  | 10,631,364 | 11,573,230  |
| 5 to 14              | 16,954,357 | 18,867,772 | 22,039,212  |
| 15 to 19             | 7,556,089  | 9,063,603  | 9,430,556   |
| 20 to 44             | 28,632,443 | 35,866,859 | 40,555,543  |
| 45 and over          | 13,480,474 | 17,373,613 | 21,963,380  |
| Age unknown          | 200,584    | 169,055    | 148,699     |
| Total                | 75,994,575 | 91,972,266 | 105,710,620 |

Another important arrangement of data into classes or groups is by means of *quantitative* classification, or classi-

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 780, Table 759.

<sup>2</sup> *Statistical Abstract of the United States, 1928*, page 5, Table 9.



fication according to size of some inherent factor in relation to some other numerical factor. By this method of classification the data are merely divided into groups according to particular characteristics that can be numerically expressed. The preceding illustrative table will show the exact nature of this kind of classification.

The data in the preceding table are in reality a frequency distribution, as is any other classification made according to the number of items or individuals of specified size that belong to a class of designated size or quality. In Table VI, for instance, it is shown that there were in 1920, 11,573,230 children in the United States below the age of 5 years. The age class is 5 years, and the size of the group belonging to this class is expressed by the number 11,573,230. Each is dependent upon the other. If the class were greater than 5 years, then the number of individuals belonging to it would be greater than 11,573,230, for it would include children who are more than 5 years old. Conversely, if the number were greater than 11,573,230 the class would necessarily be greater than 5, assuming the same distribution as referred to in Table VI. A further discussion of this type of classification will be found in Chapter III.

There is, then, nothing difficult in regard to classification of data, since all groups of whatever nature readily lend themselves to one or the other of the classifications that have been described. The individuals constituting a class can always be grouped on the basis of either quality, location, or quantity.

### CONSTRUCTION OF TABLES

In considering the construction of statistical tables it is well to appreciate the fact that tables for presentation

purposes and tables for office use may be somewhat unlike, and yet convey the same information. For office use the titles, headings, captions, and footnotes may be abbreviated, subheadings often omitted, and columns and totals of columns need not necessarily be separated by rulings. To facilitate reading of figures that are not compiled for presentation purposes spacings may be left between rows to enable one to more readily comprehend magnitudes, relationships, and relative positions. There are many other variations from the form of preparing tables for presentation that may be observed in office work, and observance of these variations often prevents the waste of much valuable time. It is only when tables are intended for general distribution that any great amount of time should be devoted to descriptive and explanatory factors. Since there is an ever-increasing tendency to present information in concise form, eliminating long and tiresome explanations, and to employ statistical tables very extensively, a brief discussion will be given to the preparation of tables for this purpose.

To present data in such an arrangement as to enable the reader to grasp the meaning and significance of the figures immediately, or as easily as possible, is one of the most important phases of statistical technique. It is not sufficient that a table merely show the mass of figures involved. They must be so arranged in columns and under appropriate descriptive headings, or captions, that their relation to the title of the table and to the subject in general is clearly understood. To this end, every table should be complete within itself, containing ample explanation in regard to the meaning of the items or figures shown. The following table may be taken as an example of the simplest sort.

TABLE VII. POPULATION OF CONTINENTAL UNITED STATES, 1924-1928 <sup>1</sup>

| YEAR | NUMBER      |
|------|-------------|
| 1924 | 113,727,432 |
| 1925 | 115,378,094 |
| 1926 | 117,136,000 |
| 1927 | 118,628,000 |
| 1928 | 120,013,000 |

The reader immediately understands this table. Its title and heading are self-explanatory, and there is no doubt as to the meaning of the figures. Furthermore, the title of the table and the heading of the column are brief and unambiguous, with the customary footnote giving the source of the data.

Figures in tables of this general kind where large numbers are involved are often expressed in contracted form, as shown below in Table VIII.

TABLE VIII. PRODUCTION OF APPLES IN THE UNITED STATES, 1924-1928 <sup>2</sup>

| YEAR | PRODUCTION           |
|------|----------------------|
|      | <i>1,000 bushels</i> |
| 1924 | 171,725              |
| 1925 | 172,389              |
| 1926 | 246,524              |
| 1927 | 123,693              |
| 1928 | 184,920              |

Whenever a table is made up of long columns of figures, it is a good plan to leave a space at regular intervals to facilitate reading and to avoid confusion. This interval may be left after each five or ten figures, depending upon the nature of the data, the length of the columns, the size of the figures, and personal preference. The following table will illustrate this principle.

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 3, Table 6.

<sup>2</sup> *Yearbook, United States Department of Agriculture, 1928*, page 764, Table 128.

TABLE IX. CAPITAL ISSUES OF CORPORATIONS IN THE UNITED STATES <sup>1</sup>

| YEAR | TOTAL CAPITAL ISSUES |
|------|----------------------|
|      | <i>1,000 dollars</i> |
| 1907 | 1,393,913            |
| 1908 | 1,423,199            |
| 1909 | 1,681,621            |
| 1910 | 1,518,272            |
| 1911 | 1,739,488            |
| 1912 | 2,253,587            |
| 1913 | 1,645,736            |
| 1914 | 1,436,518            |
| 1915 | 1,435,351            |
| 1916 | 2,186,500            |
| 1917 | 1,529,970            |
| 1918 | 1,344,810            |
| 1919 | 3,021,171            |
| 1920 | 3,106,931            |
| 1921 | 2,634,869            |
| 1922 | 3,423,948            |
| 1923 | 3,601,438            |
| 1924 | 3,219,146            |
| 1925 | 3,642,012            |
| 1926 | 3,643,646            |
| 1927 | 5,970,297            |

In a report dealing with one subject or in which the same subject occupies considerable space, the contents of tables are often made more easy to grasp by repeatedly using the name of that subject as the preliminary to every title. The United States Department of Agriculture is one of the federal agencies using this method extensively. On page 39 is a table to illustrate this simplicity in titles.

The Maryland Agricultural Experiment Station reported experiments with different kinds of lime, covering eleven years, with a rotation of corn, wheat, and hay

<sup>1</sup> *Statistical Abstract of the United States, 1928, page 309, Table 321.*

TABLE X. COTTON: YIELD OF LINT PER ACRE IN SPECIFIED COUNTRIES, CROP YEAR OF 1927-28 <sup>1</sup>

| COUNTRY          | YIELD PER ACRE |
|------------------|----------------|
|                  | <i>Pounds</i>  |
| United States    | 155            |
| India            | 95             |
| Egypt            | 382            |
| Brazil           | 194            |
| Russia (Asiatic) | 182            |
| Mexico           | 263            |
| Chosen (Korea)   | 126            |
| Argentina        | 231            |

(timothy and clover), 1,400 pounds of calcium oxid (burned lime) and equivalent amounts of calcium carbonate (ground oyster shells and shell marl) having been applied per acre at the beginning. Four crops of corn, three of wheat, and four of hay were harvested during the eleven years. In commenting on the results, Director Patterson says: "It will be noted that the carbonate of lime gave decidedly better results than the caustic lime." <sup>2</sup>

TABLE XI. EXPERIMENTS WITH LIME IN MARYLAND <sup>3</sup>

| KIND OF LIME USED<br>IN EXPERIMENTS   | TOTAL PRODUCTION OF SPECIFIED CROPS IN<br>ELEVEN YEARS |                 |                |
|---------------------------------------|--------------------------------------------------------|-----------------|----------------|
|                                       | Hay (4 crops)                                          | Wheat (3 crops) | Corn (4 crops) |
|                                       | <i>Tons</i>                                            | <i>Bushels</i>  | <i>Bushels</i> |
| None                                  | 2.60                                                   | 32              | 98             |
| Caustic lime burned<br>from stone     | 3.09                                                   | 32              | 128            |
| Caustic lime burned<br>from shells    | 3.82                                                   | 34              | 129            |
| Calcium carbonate in<br>ground shells | 3.92                                                   | 42              | 148            |
| Calcium carbonate in<br>shell marl    | 4.29                                                   | 43              | 145            |

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 840, Table 256.

<sup>2</sup> Soil Fertility and Permanent Agriculture, C. G. Hopkins, pages 167-168.

<sup>3</sup> Maryland Agricultural Experiment Station records, and Soil Fertility and Permanent Agriculture, C. G. Hopkins, page 167.

With these results at hand the problem is to present them in such a form as to be easily understood. The preceding table illustrates a satisfactory way by which they may be shown.

The United States Tariff Commission employs a very satisfactory method of presenting a large amount of statistical data in compact, orderly, and logical form. One of the tables prepared by the Commission's statistical clerks is presented on page 41.

So many factors enter into the construction of tables, such as nature of the data, purpose of the table, amount of information to be shown, and individual preferences, that no attempt can be made to show a very large number of the many tabular forms. No one can be looked upon as a final authority on the technique of table construction, for every classification, tabulation, or other arrangement of data has its own problems. General usage and continued experience, however, have evolved certain rather definite rules and principles that may well be observed by the beginner. A few of these will be briefly summarized.

1. All tables should be complete within themselves. They should contain explanations, such as column headings, that show clearly the exact meaning of the data. These headings should be as brief as possible.

2. The title and number of a table should always be at the top, and the title itself should be concise, and yet lengthy enough so that the reader will be able to gain a clear conception of the table's contents without finding it necessary to refer to the text for explanations. Whenever a table is presented expressly for the purpose of showing a certain series of data for a specific place and for a definite period of time it should embody answers to the queries "What," "Where," and "When."



TABLE XII. IMPORTS OF WHEELS FOR RAILWAY PURPOSES <sup>1</sup>

| CALENDAR YEAR              | RATE OF DUTY     | QUANTITY                 | VALUE   | DUTY COLLECTED | VALUE PER UNIT OF QUANTITY | ACTUAL OR COMPUTED AD VALOREM RATE OF DUTY |
|----------------------------|------------------|--------------------------|---------|----------------|----------------------------|--------------------------------------------|
| 1920                       | 20 per cent.     | <i>Pounds</i><br>258,360 | \$8,392 | \$1,678        | \$0.032                    | <i>Per Cent.</i><br>20.00                  |
| 1921                       | "                | 67,717                   | 1,496   | 299            | .022                       | 20.00                                      |
| 1922 (Jan. 1-<br>Sept. 21) | "                | 14,015                   | 939     | 188            | .067                       | 20.00                                      |
| 1922 (Sept.<br>22-Dec. 31) | 1 cent per pound | 13,274                   | 436     | 133            | .033                       | 30.44                                      |
| 1923                       | "                | 766,058                  | 92,215  | 7,661          | .120                       | 8.31                                       |
| 1924                       | "                | 720,671                  | 28,983  | 7,207          | .040                       | 24.87                                      |
| 1925                       | "                | 2,144,230                | 86,169  | 21,442         | .040                       | 24.88                                      |
| 1926                       | "                | 1,034,571                | 41,614  | 10,346         | .040                       | 24.86                                      |
| 1927                       | "                | 2,082,188                | 84,460  | 20,822         | .041                       | 24.65                                      |
| 1928                       | "                | 2,438,612                | 123,188 | 24,386         | .051                       | 19.80                                      |

<sup>1</sup> Summary of Tariff Information, 1929, Schedule 3, page 690.

3. Footnotes should be used to show sources of data and to call attention to other factors of importance which cannot be included in the body of the table.

4. A table should be so constructed as to include only such data as are similar or very closely related. Table XII furnishes an example of factors that bear close and significant relationship.

5. Columns of figures should be separated by lines of such shade and width as to make it easy for the reader to read them.

6. Tables should be so constructed that the several columns of figures will occupy space in proportion to size and relative importance.

7. The units in which figures in a column are expressed should be clearly indicated, and they should be identical when one column is being compared with another.

8. Whenever a column heading has relation to several closely associated factors the boxing which incloses the heading should be modified to show this. For example, we may have in one table the total value of strawberries, the value per crate, and the value per quart. Obviously, if the columns showing these values are adjoining they may be placed under the general heading of "Value," with subheadings designated to show total value, value per crate, and value per quart.

9. Letters or numbers, in addition to column headings, or other kind of captions, should be given to columns and rows when it is desired to make frequent reference in the text to certain parts of tables.

10. In rows of figures, indentation is used to set off or call attention to relationships. If, for instance, we were constructing a table to show the urban and rural population of the country for three decades by geo-

graphical divisions and by states, we could indent the names of the states under the name of the division to show that the state is subordinate. A further aid in showing this relationship would be to print the name of the division in heavier type.

11. Columns and rows of figures that are to be expressed as totals should be set off by either a space or a line. The construction of the particular table will determine which is to be used.

12. As has been shown, if a table is exceptionally long, a space should be used to separate about every fifth row of figures, or every group of five figures when the table consists of a single column.

13. Whenever some items are missing, if the number is not too great, it is a good plan to leave a vacant space to show this break. For instance, for the period from 1920 to 1930 the data on production of steel rails of a certain weight may be available for all years except 1925, but in constructing a table it would be a wise plan to include 1925 in the table and leave a vacant space in the column to show that the production for that year is unknown. Foot-note references in such instances as this are advisable. A dash placed in the column opposite 1925 might be taken to indicate there was no production that year.

14. Commas and decimal points in a column must be kept in alignment.

15. If practicable, either the data in the columns or the designations given to rows of figures should be arranged in some consecutive order. Let us say that grades received by a group of 50 pupils are being recorded in a single column. In this instance a good plan would be to arrange them in either ascending or descending order of magnitude. Again, if a table were being made to show

the production of steel ingots and castings for a 10-year period, the years would be arranged in consecutive order without any regard to size or quantity of production, if the aim were merely to show the fluctuations from year to year.

16. Large numbers should be separated by commas, and dollars and cents, as well as all other numbers in which fractional parts are involved, should be clearly separated by decimal points.

17. When a report is being prepared the tables should be placed as near as possible to the text material that relates to them. Circumstances will determine whether the table is to precede or to follow the text. There are times when it is easier to place the table early in the text and then follow it with a description or interpretation, and there are other times when it is better for the table to come after the explanatory discussion. These are problems that must be decided on the basis of their own merit.

18. Heavy type may be used to call attention to particular items in a column or row of figures. A table may be constructed to show the domestic production of all makes of automobiles, and when completed it may be desired to place emphasis, say, upon the factory output of Chryslers. This could be done by printing the name and the production figures in heavy type. The same effect could be brought about, of course, by placing the production figures at either the top or bottom of the table and leaving a space between them and the next items, or they could be placed in the middle of the table and a space left on either side.

19. The text is improved by using larger type, and often heavier type, for all table titles.

## CHAPTER III

### FREQUENCY DISTRIBUTIONS

In the interpretation of many statistical data the first problem is to classify the items into groups according to size or value, or some other attribute. We shall find it to our advantage in considering the technique of frequency distributions to acquaint ourselves with the meaning of several important terms in common usage among statisticians. The first of these is the *variable*, which is simply a number, symbol, or some other characteristic that is used to show the size or attribute of the items constituting a group or mass of data. The following table will give a clear conception of what a variable is.

TABLE XIII. GRADES MADE BY TWENTY-EIGHT SIXTH-GRADE PUPILS  
ON A GEOGRAPHY TEST <sup>1</sup>

| GRADE | NUMBER OF PUPILS RECEIVING<br>THE GRADE<br>(Frequency) |
|-------|--------------------------------------------------------|
| 88    | 4                                                      |
| 89    | 4                                                      |
| 90    | 7                                                      |
| 91    | 7                                                      |
| 92    | 6                                                      |
| Total | 28                                                     |

The variables in the preceding table are the grades 88, 89, 90, 91, and 92. These grades are numbers used to show the specific attainments of a group of pupils, or to measure the value of particular work done by them. All were not of

<sup>1</sup> Selected from school records in Maryland.

equal ability. There were variations in accomplishments, which are expressed by different or varying grades, ranging from 88 to 92. Therefore, these grades are the variable numerical quantities or characteristics expressing these differences, and it is in this sense that they become known as *variables*.

There are two distinct kinds of variables, known as *continuous* and *discrete*. As is implied, a *continuous* variable is one that may have any number of values ranging between the lowest and the highest, such as the weight of human beings, farm costs of production, height of trees, etc. The *discrete* variable, on the other hand, is one with a value distinct and separate, as, for example, the number of professors at an institution, or the number of pupils in an elementary school. Such numbers cannot occur in unbroken continuity, for there can be no fractional parts.

The second term to consider is *variate value*. This is merely the value presented by the variable, or by the characteristic, having more than one numerical value. In Table XIII the varying magnitudes or sizes of the grades are the variate values.

TABLE XIV. GROSS EARNINGS OF FEDERAL RESERVE BANKS, 1918-1927 <sup>1</sup>

| YEAR | GROSS EARNINGS       |
|------|----------------------|
|      | <i>1,000 dollars</i> |
| 1918 | 67,584               |
| 1919 | 102,381              |
| 1920 | 181,297              |
| 1921 | 122,866              |
| 1922 | 50,499               |
| 1923 | 50,709               |
| 1924 | 38,340               |
| 1925 | 41,801               |
| 1926 | 47,600               |
| 1927 | 43,024               |

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 251, Table 255.



A *series* is the total number of values arranged in some definite order and showing the differences between items. An illustration is presented in the preceding table, in which the magnitudes of gross earnings constitute the series.

The arrangement of the items, the gross earnings, in the table above is according to orderly sequence, and in this way they show the general nature of changes in these earnings from year to year for the entire period.

Turning to Table XIII we recall that the numbers representing grades, 88, 89, 90, 91, and 92 are variables. This will serve as an aid in comprehending the meaning of *variate frequency*, which is the number showing how many times a variable occurs. Accordingly, in Table XIII the *variate frequency* for the variable 88 is 4, meaning, of course, that the grade or value of 88 occurs 4 times.

The next term with which to acquaint ourselves is *class frequency*. Its meaning will be immediately grasped after reading the table below.

TABLE XV. GRADES MADE BY THIRTY-NINE SIXTH-GRADE PUPILS ON A SCIENCE TEST <sup>1</sup>

| GRADE<br>(Class) | NUMBER OF PUPILS RECEIVING<br>THE GRADE<br>(Frequency) |
|------------------|--------------------------------------------------------|
| 81-85            | 8                                                      |
| 86-90            | 9                                                      |
| 91-95            | 10                                                     |
| 96-100           | 12                                                     |
| Total            | 39                                                     |

It will be seen that the grades have been grouped into classes. Let us take the last class, which is 96-100, including all grades from 96 to 100, inclusive. The number of pupils receiving grades that fall within that class is 12.

<sup>1</sup> Selected from school records in Maryland.

Therefore, 12 is the *class frequency*, since it shows how many times the grades within that class occur.

The *frequency distribution* is the apportionment (or distribution) of the items over the variate values. Thus in Table XV the number of pupils, 8, 9, 10, and 12, considered collectively, constitutes the frequency distribution. Likewise, in Table XIII the numbers 4, 4, 7, 7, and 6 are also a frequency distribution. It will be seen, then, that a *frequency table*, the next term to be defined, is simply a table showing a frequency distribution.

A *frequency surface* represents the frequency distribution for two variables, and is, therefore, described by a surface or solid. This surface is in fact merely a graphic representation of the data in a correlation table.<sup>1</sup> If perpendiculars are erected at the center of each compartment, of lengths proportionate to the frequency in that compartment, the ends of these perpendiculars will represent points on the surface.

There are four other common statistical terms which should be understood early by the beginner. These will be explained by referring again to Table XV. The first is *class interval*. Let us take the class 81–85 to illustrate the meaning. We see that the class extends from 81 to 85, inclusive. In other words, the interval of the class is 5, since to move or ascend from 81 to 85 a count of 5 is required, and it may be defined as the interval that limits the size or extent of a class in a frequency distribution. In Table XIII the class interval is 1.

A discussion of the class interval naturally lends itself to what its magnitude should be. This is a decision that must be reached on the basis of the nature of the data

<sup>1</sup> Table LXXVIII, page 237, Chapter XII, illustrates the construction of a correlation table.

themselves in regard to size and number of items and the purpose for which used or intended. In general, the class interval should be small when there is a small number of items and large when the number of items is large. While the class interval should be uniform throughout when possible to have it so, there are frequent instances when the lowest and highest values in a distribution are so scattered as to make it impracticable to arrange them in accordance with the general scheme of classification. The following table is constructed on such a basis.

TABLE XVI. FREQUENCY DISTRIBUTION OF COTTON REPORTS, BY YIELD GROUPS, 1927 <sup>1</sup>

| YIELD GROUP<br>(Pounds of lint per acre)<br>(Class) | NUMBER OF REPORTS<br>(Frequency) |
|-----------------------------------------------------|----------------------------------|
| 60 and under                                        | 45                               |
| 61 to 100                                           | 72                               |
| 101 to 140                                          | 90                               |
| 141 to 180                                          | 135                              |
| 181 to 220                                          | 117                              |
| 221 to 260                                          | 197                              |
| 261 to 300                                          | 102                              |
| 301 to 340                                          | 32                               |
| 341 to 380                                          | 54                               |
| 381 to 420                                          | 52                               |
| 421 to 460                                          | 26                               |
| 461 to 500                                          | 47                               |
| 501 and over                                        | 23                               |
| Total                                               | 992                              |

In the preceding table the class interval is 40 with the exception of the first yield group, which shows all the reports with yields per acre of 60 pounds and less, and the last yield group, showing the number of reports for yield per acre of 501 pounds and more.

The *class limits* are the upper and lower extremes of a class. In Table XV, for example, taking the first class,

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 1049, Table 521.

81-85, the lower limit as expressed is 81 and the expressed upper limit is 85. In practice it may be necessary to recognize a distinction between the expressed class limits and the actual class limits. The expressed limits of the group 81-85, as we have seen, are 81 and 85, but actually the limits are 80.50 and 85.499+. That is to say, any value between 80.50 and 85.499+ belongs to the 81-85 class.

The midpoint or midvalue of a class is sometimes referred to as the *class mark*. In Table XV, then, the class marks of the several classes are 83, 88, 93, and 98, these being the midvalues or midpoints of the respective classes.

An *array* is the arrangement of items in order of ascending or descending magnitude. It differs from a *series* in that the latter is not necessarily an arrangement on the basis of magnitudes, but is rather an arrangement in logical sequence, as in Table XIV. An array may be known also as an *order distribution*. In the following table is an array of school grades arranged in ascending order of magnitude.

TABLE XVII. ARRAY OF GRADES MADE BY FORTY SIXTH-GRADE PUPILS ON A SPELLING TEST <sup>1</sup>

|    |    |    |    |     |
|----|----|----|----|-----|
| 80 | 85 | 90 | 94 | 98  |
| 80 | 87 | 90 | 94 | 99  |
| 81 | 87 | 91 | 95 | 99  |
| 82 | 87 | 92 | 95 | 99  |
| 82 | 88 | 93 | 95 | 100 |
| 83 | 88 | 93 | 96 | 100 |
| 84 | 89 | 93 | 97 | 100 |
| 84 | 90 | 93 | 97 | 100 |

In making an array of items of which two or more are of the same value or magnitude it is a good plan to ar-

<sup>1</sup> Selected from school records in Maryland.

## FREQUENCY DISTRIBUTIONS

range these items by some definite scheme. Identical grades in the preceding array are arranged according to the spelling of names. In the case of grade 97, for example, which occurs twice, the grade received by a pupil named Jones would occur in the array before the grade made by Williams.

The *range* is the spread between the smallest and largest value or item in an array or other arrangement of data. If the smallest value is 81 and the largest value is 90, then the range, or spread, between the two values is 10. The range, then, in such an instance is merely the difference between the lower and upper limits of the extremes.

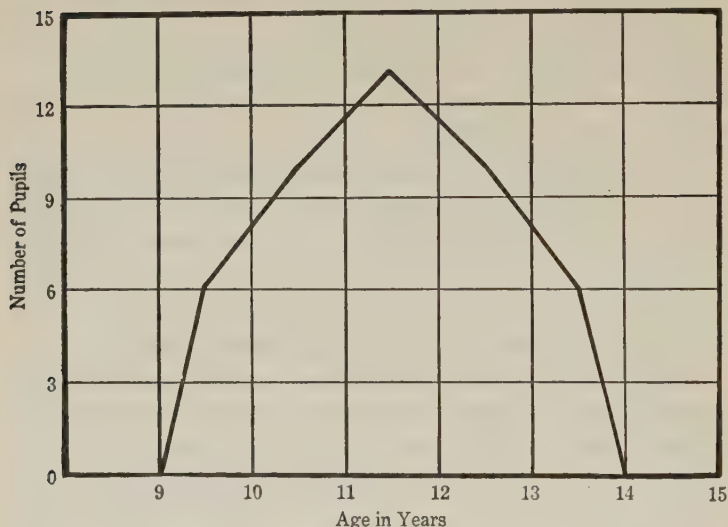
Statistical data lend themselves to two general classifications, *ungrouped* and *grouped*. These two terms are self-explanatory. In Table XVII, for example, we have an ungrouped array of grades made by pupils on a spelling test. These grades are simply recorded in their order of magnitude without any attempt having been made to classify them into groups. Table XVIII shows a set of ungrouped data which are not arrayed.

Grouped data are represented by the frequency distribution in Table XIX. Here we have the ungrouped data in Table XVIII arranged according to classes.

## SIMPLE DISTRIBUTION

A simple distribution is merely the arrangement or disposition of individuals according to some definite classification. Let us assume we have a group of 45 grammar school pupils ranging in age from 10 to 14 years. The classification of these 45 pupils into smaller groups according to age would constitute a simple frequency distribution. Chart I illustrates one method of presenting graphically such a distribution.

CHART 1. DISTRIBUTION OF PUPILS ACCORDING TO AGE  
(Based on assumed data)



In the following table we have the grades of forty-seven pupils recorded in the identical order in which the names of the pupils appeared on the teacher's alphabetically arranged roll.

TABLE XVIII. GRADES MADE BY FORTY-SEVEN SIXTH-GRADE PUPILS ON A MATHEMATICS TEST <sup>1</sup>

|    |     |    |     |
|----|-----|----|-----|
| 98 | 89  | 91 | 87  |
| 97 | 100 | 74 | 99  |
| 75 | 99  | 84 | 98  |
| 87 | 73  | 83 | 86  |
| 88 | 88  | 82 | 87  |
| 92 | 89  | 89 | 86  |
| 91 | 76  | 95 | 88  |
| 85 | 94  | 93 | 78  |
| 84 | 71  | 77 | 93  |
| 80 | 85  | 83 | 95  |
| 79 | 76  | 82 | 100 |
| 77 | 81  | 83 |     |

<sup>1</sup> Selected from school records in Maryland.



In this arrangement it is rather difficult to comprehend the distribution of grades, since they are not arrayed according to magnitudes. A frequency table corrects this condition for us. Let us observe the following table.

TABLE XIX. FREQUENCY DISTRIBUTION OF GRADES MADE BY FORTY-SEVEN SIXTH-GRADE PUPILS ON A MATHEMATICS TEST <sup>1</sup>

| GRADE<br>(Class) | NUMBER OF PUPILS RECEIVING<br>SPECIFIED GRADE<br>(Frequency) |
|------------------|--------------------------------------------------------------|
| 71-75            | 4                                                            |
| 76-80            | 7                                                            |
| 81-85            | 10                                                           |
| 86-90            | 11                                                           |
| 91-95            | 8                                                            |
| 96-100           | 7                                                            |
| Total            | 47                                                           |

We can immediately see after arrangement of the data in a frequency table how many pupils received grades within specific limits. This is a simple frequency distribution, meaning that the frequencies recorded for each class include only those belonging to that particular class. In other words, the number of items in a specific class is separate from the number of all other items of other classes.

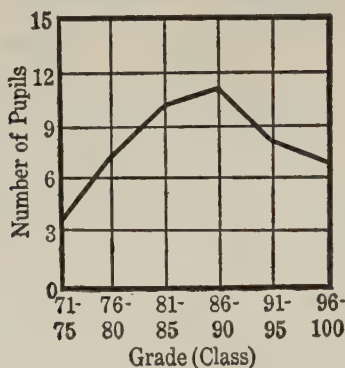
If the number of pupils were large enough we could construct a frequency table showing the number receiving each grade, so that instead of the classes, 71-75, 76-80, etc., we would have each individual grade from 71 to 100, with the number of pupils receiving each grade recorded in the frequency column.

The data in Table XIX are presented graphically in Chart 2.

<sup>1</sup> Data from Table XVIII, page 52.

CHART 2. GRADES MADE BY FORTY-SEVEN SIXTH-GRADE PUPILS ON A MATHEMATICS TEST

(Based on data in Table XIX)



## CUMULATIVE DISTRIBUTION

When a frequency table is so constructed that the frequencies recorded opposite a class include not only those belonging to that particular class, but also those belonging to preceding classes, we have what is known as a cumulative frequency distribution. This is because the frequencies are cumulated or superadded. The following table illustrates this kind of frequency.

TABLE XX. CUMULATIVE FREQUENCY DISTRIBUTION OF GRADES MADE BY FORTY-SEVEN SIXTH-GRADE PUPILS ON A MATHEMATICS TEST <sup>1</sup>

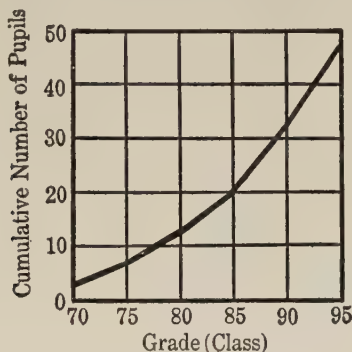
| GRADE<br>(Class) | SIMPLE<br>FREQUENCY | CUMULATIVE<br>FREQUENCY |
|------------------|---------------------|-------------------------|
| 70               | 3                   | 3                       |
| 75               | 4                   | 7                       |
| 80               | 6                   | 13                      |
| 85               | 7                   | 20                      |
| 90               | 12                  | 32                      |
| 95               | 15                  | 47                      |

In Chart 3 there is presented graphically the cumulative frequency distribution shown in Table XX.

<sup>1</sup> Based on data specially selected from Maryland school records.

CHART 3. CUMULATIVE DISTRIBUTION OF GRADES MADE BY FORTY-SEVEN SIXTH-GRADE PUPILS ON A MATHEMATICS TEST

(Based on data in Table XX)



## SPATIAL DISTRIBUTION

As is implied, a spatial distribution is based on space or geographical and linear distance. We may wish to show, for example, how many farmers living at different distances from town own motor trucks, how many employees in an office building live within specified distances of their work, the number of gasoline filling stations along a high-

TABLE XXI. NUMBER OF PUPILS LIVING WITHIN SPECIFIED DISTANCES OF THE SCHOOL THEY ATTEND <sup>1</sup>

| DISTANCE FROM SCHOOL<br>AS MEASURED BY NUMBER<br>OF CITY BLOCKS<br>(Class) | NUMBER OF PUPILS LIVING<br>WITHIN SPECIFIED DISTANCES<br>OF THE SCHOOL<br>(Frequency) |
|----------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 0-1.0                                                                      | 3                                                                                     |
| 1.1-2.0                                                                    | 7                                                                                     |
| 2.1-3.0                                                                    | 7                                                                                     |
| 3.1-4.0                                                                    | 9                                                                                     |
| 4.1-5.0                                                                    | 5                                                                                     |
| 5.1-6.0                                                                    | 5                                                                                     |
| 6.1-7.0                                                                    | 4                                                                                     |
| Total                                                                      | 40                                                                                    |

<sup>1</sup> Specially selected data from school records in Baltimore City.

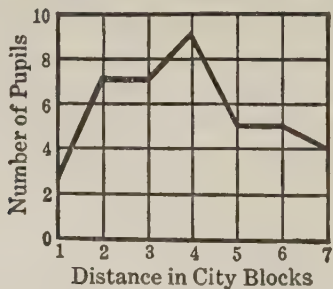
way and their distances from a fixed point, or the quantity of fruit that is produced in localities of various distances from the packing plant. The table on page 55 will show the general nature of a spatial distribution.

This table shows at a glance how many of the 40 pupils live at certain specified distances from the school they attend. The number of pupils may be said to be grouped or classified, therefore, according to where they live in the space or area around the school. This is one of the simplest examples of spatial distributions.

A graphic presentation of the data in Table XXI is shown in Chart 4.

CHART 4. DISTRIBUTION OF PUPILS ACCORDING TO DISTANCES THEY LIVE FROM SCHOOL

(Based on data in Table XXI. The ordinates of the line are purposely plotted directly on the x values rather than midway between them.)



A graphic presentation of spatial distributions may also be made by the use of maps, in which the frequencies in each location or area are represented by series of dots rather than by a single dot for each area and varied in size according to the relative importance of each area. Let us make this clear by means of an illustration. It is desired to show graphically, for example, the location of poultry by counties on farms in North Carolina. There are two common ways by which this may be done. In

1926 there were approximately 111,000 laying hens on farms in Wake County.<sup>1</sup> This number may be represented by a single dot of, say, one-fourth of a square inch in area. Then for each of all the other counties in the state one dot would be used, with the area or size of each dot bearing the same ratio to the one used for Wake County as the ratio between the number of laying hens in that county and each of the other counties. If the number of laying hens in a county in eastern Carolina is twice as great as the number in Wake County, then the single dot used to represent that number would be twice as great in area as the dot used to represent the number in Wake County. This is the general nature of a *single dot* map. The one great disadvantage of maps of this kind is that the reader cannot form a true estimate of the difference in the size of the various dots, and for that reason is often not able to clearly interpret them.

A more satisfactory way to present a spatial distribution graphically is by means of a map in which several or more dots are used to show the location of items. Let us revert to the example in the preceding paragraph. There were, we have seen, approximately 111,000 laying hens in Wake County in 1926. The best way to show this figure on a map is by a number or series of dots. We may let one dot represent 10,000, in which case 11 dots and a fraction would be used to represent the total of 111,000. The fractional part of the dot would be a shaded portion extending over one-tenth of the area of a circle of the same size in area as the completely shaded ones, with the remaining portion of the circle unshaded. The commendable feature of the *multiple dot* map is that each complete

<sup>1</sup> The exact number as reported by the Farm Census of North Carolina is 110,911.

dot represents the same number of items, so that the reader can more clearly estimate the totals for the various areas. In this instance all that one needs to do in ascertaining the total is to count the number of dots or circles and multiply by the number of items that each dot or circle represents, whereas when the *single dot* map is used an estimate must be made of the relative areas of the various dots, which is a very difficult thing to do. The *multiple dot* map corrects this condition, and constitutes probably the most satisfactory of all common methods of graphically presenting a spatial distribution.

### TEMPORAL DISTRIBUTION

A temporal distribution is the arrangement of frequencies in the order of time in which they occur. This kind of distribution divides a total number of items belonging to a certain period of time into smaller parts

TABLE XXII. IMPORTS OF SHELLED PEANUTS FOR CONSUMPTION, BY MONTHS, 1927 <sup>1</sup>

| MONTH     | IMPORTS FOR CONSUMPTION |
|-----------|-------------------------|
|           | <i>1,000 pounds</i>     |
| January   | 2,724                   |
| February  | 3,503                   |
| March     | 4,111                   |
| April     | 4,652                   |
| May       | 2,690                   |
| June      | 2,571                   |
| July      | 3,144                   |
| August    | 2,313                   |
| September | 2,307                   |
| October   | 1,687                   |
| November  | 1,265                   |
| December  | 486                     |
| Total     | 31,453                  |

<sup>1</sup> *Peanuts, Report of the United States Tariff Commission to the President of the United States, 1929, Appendix C, page 86, Table 75.*



belonging to shorter periods. Temporal distributions will be illustrated in several forms. The first will be a monthly distribution.

The preceding table shows the total imports of shelled peanuts for the entire year of 1927, and in addition it shows the distribution of these imports by months. In other words, the total receipts, 31,453,000 pounds, for the year are divided into 12 parts on the basis of the time they entered the ports of this country. This, then, is a temporal, or time, distribution.

CHART 5. IMPORTS OF SHELLED PEANUTS FOR CONSUMPTION, BY MONTHS, 1927

(Based on data in Table XXII)

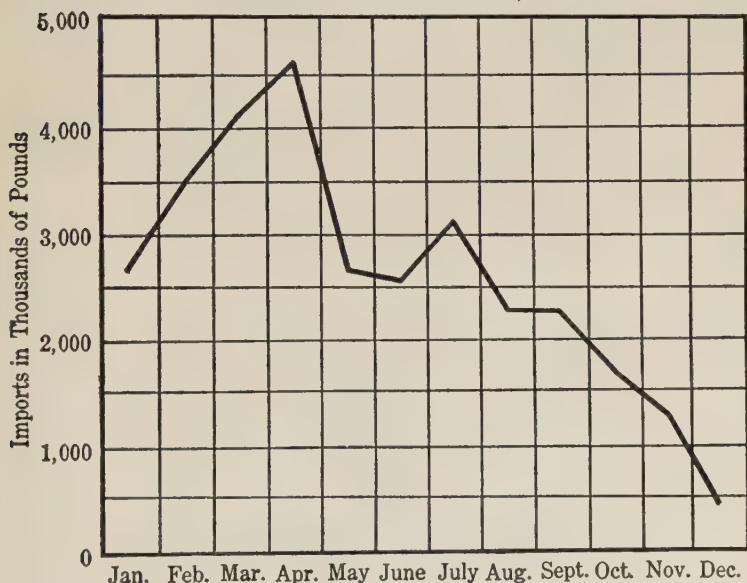


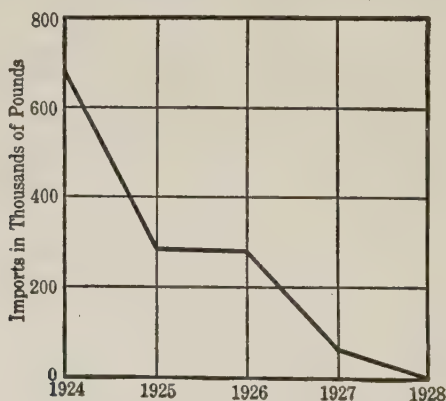
Chart 5 shows graphically the distribution of these imports throughout the year.

TABLE XXIII. IMPORTS OF CITRIC ACID, 1924-1928 <sup>1</sup>

| CALENDAR YEAR | QUANTITY OF IMPORTS |
|---------------|---------------------|
|               | <i>Pounds</i>       |
| 1924          | 673,114             |
| 1925          | 288,574             |
| 1926          | 284,897             |
| 1927          | 71,291              |
| 1928          | 1,338               |
| Total         | 1,319,214           |

Table XXIII shows a temporal distribution by years. From 1924 to 1928, inclusive, the imports of citric acid into the United States aggregated 1,319,214 pounds, but we wish to know how these imports were distributed over the five-year period. The above table presents not only the statistics of the total for this period, but also the statistics of imports for each year separately.

CHART 6. IMPORTS OF CITRIC ACID INTO THE UNITED STATES, 1924-1928  
(Based on data in Table XXIII)



A graphic presentation of the imports of citric acid as shown in Table XXIII is given in Chart 6.

<sup>1</sup> Summary of Tariff Information, 1929, Schedule 1, page 7.

For long periods of time it may be advisable to present the data as aggregates for periods of time of uniform length, particularly when interest lies only in the general course of events. The following table shows the total number of immigrants entering the United States during the period from 1871 to 1920, inclusive. It shows also the number of immigrants by decades. This is merely another type of temporal distribution, with aggregates by decades being presented instead of the data for each individual year.

TABLE XXIV. IMMIGRATION FROM ALL COUNTRIES INTO THE UNITED STATES FOR FIVE DECADES, 1871-1880 TO 1911-1920 <sup>1</sup>

| DECADE    | TOTAL IMMIGRANTS |
|-----------|------------------|
|           | <i>Number</i>    |
| 1871-1880 | 2,812,191        |
| 1881-1890 | 5,246,613        |
| 1891-1900 | 3,687,564        |
| 1901-1910 | 8,795,386        |
| 1911-1920 | 5,735,811        |
| Total     | 26,277,565       |

### SYMMETRICAL DISTRIBUTION

A distribution is symmetrical when the number of items or frequencies occurring at the same linear distance on either side of the midpoint of the frequencies is identical. It has been defined as "an arrangement of pairs of points in a general system such that the set of lines joining them together is divided into equal parts by a line (the axis of symmetry), or by a plane (plane of symmetry), or by a point (the center of symmetry)." The following tables are presented to show the more exact nature, and also the varying nature, of this type of distribution.

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 93, Table 101.

TABLE XXV. SYMMETRICAL DISTRIBUTION OF GRADES MADE BY SIXTY-FOUR SIXTH-GRADE PUPILS ON A VOCABULARY TEST <sup>1</sup>

| GRADE<br>(Class) | NUMBER OF PUPILS RECEIVING<br>THE GRADE<br>(Frequency) |
|------------------|--------------------------------------------------------|
| 65               | 1                                                      |
| 70               | 6                                                      |
| 75               | 15                                                     |
| 80               | 20                                                     |
| 85               | 15                                                     |
| 90               | 6                                                      |
| 95               | 1                                                      |
| Total            | 64                                                     |

TABLE XXVI. SYMMETRICAL DISTRIBUTION OF HEIGHTS OF CORNSTALKS <sup>1</sup>

| HEIGHT IN FEET<br>(Class) | NUMBER OF INDIVIDUALS OF<br>SPECIFIED HEIGHT<br>(Frequency) |
|---------------------------|-------------------------------------------------------------|
| 7.8                       | 1                                                           |
| 7.9                       | 7                                                           |
| 8.0                       | 21                                                          |
| 8.1                       | 35                                                          |
| 8.2                       | 35                                                          |
| 8.3                       | 21                                                          |
| 8.4                       | 7                                                           |
| 8.5                       | 1                                                           |
| Total                     | 128                                                         |

TABLE XXVII. SYMMETRICAL DISTRIBUTION OF HEIGHTS OF BUILDINGS <sup>1</sup>

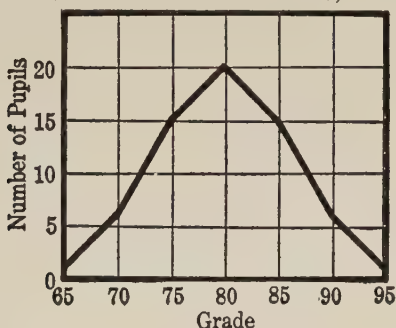
| HEIGHT OF BUILDINGS<br>IN FEET<br>(Class) | NUMBER OF BUILDINGS OF<br>SPECIFIED HEIGHT<br>(Frequency) |
|-------------------------------------------|-----------------------------------------------------------|
| 50                                        | 1                                                         |
| 75                                        | 2                                                         |
| 100                                       | 10                                                        |
| 125                                       | 2                                                         |
| 150                                       | 1                                                         |
| Total                                     | 16                                                        |

<sup>1</sup> Assumed data.

The data in Table XXV are presented graphically in Chart 7 to show more clearly the occurrence of items in one of the simplest types of symmetrical distributions.

CHART 7. SYMMETRICAL DISTRIBUTION OF GRADES MADE BY SIXTY-FOUR SIXTH-GRADE PUPILS ON A VOCABULARY TEST

(Based on data in Table XXV)



The principal characteristic to observe in regard to symmetrical distributions is that they need not be normal. All normal distributions, of course, are symmetrical, but all symmetrical distributions are not normal. As shown in the three preceding tables, when frequencies are uniformly distributed about a central point or location they are in perfect symmetry, but they do not necessarily follow the normal curve of distribution.

### ASYMMETRICAL DISTRIBUTION

An asymmetrical distribution is one in which there are unequal numbers of items lying to the left and right of the mode.<sup>1</sup> For example, if the modal value is 10, and there is one class to the left of the mode with 5 frequencies, and three classes to the right, one of which has 3 frequen-

<sup>1</sup> See Chapter VI, pages 112 to 116, for discussion of the mode.

cies, one, 2 frequencies, and the other, 1 frequency, then the curve of these frequencies would show asymmetry.<sup>1</sup>

### NORMAL DISTRIBUTION <sup>2</sup>

Whenever the number of items in successive classes increases and decreases at a rate that would be theoretically expected in an infinite number of instances, we have what is known as a normal-frequency distribution. It derives its name from the underlying mathematical principles relating to the probable occurrence of chance phenomena in accordance with the binomial expansion, and simply means that the values of a variate in a very large number of cases tend to be distributed uniformly according to mathematical law about the value that occurs the greatest number of times. The following table illustrates a normal-frequency distribution.

TABLE XXVIII. NORMAL-FREQUENCY DISTRIBUTION OF AVERAGE YEARLY GRADES MADE BY SIXTH-GRADE PUPILS <sup>3</sup>

| GRADE<br>(Class) | NUMBER OF PUPILS RECEIVING<br>SPECIFIED GRADE<br>(Frequency) |
|------------------|--------------------------------------------------------------|
| 50               | 1                                                            |
| 55               | 10                                                           |
| 60               | 45                                                           |
| 65               | 120                                                          |
| 70               | 210                                                          |
| 75               | 252                                                          |
| 80               | 210                                                          |
| 85               | 120                                                          |
| 90               | 45                                                           |
| 95               | 10                                                           |
| 100              | 1                                                            |
| Total            | 1,024                                                        |

<sup>1</sup> Asymmetry is sometimes referred to as skewness.

<sup>2</sup> See "Normal-Frequency Curve" in Chapter IX, pages 154 to 156.

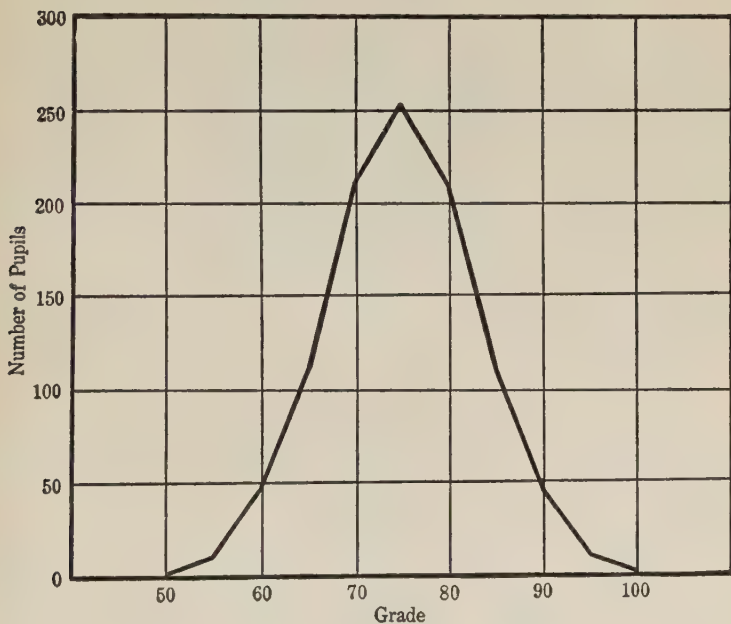
<sup>3</sup> Assumed data, constituting the binomial expansion to the tenth power.



In Chart 8 a frequency polygon is constructed which, when smoothed, shows the form of a normal-frequency curve.

CHART 8. NORMAL-FREQUENCY DISTRIBUTION OF GRADES MADE BY SIXTH-GRADE PUPILS

(Based on data in Table XXVIII)



## CHAPTER IV <sup>1</sup>

### GRAPHIC PRESENTATION

Whenever an analysis is being made of statistical data one of the most important aspects of technique is the presentation of factors in pictorial form so as to permit an immediate grasp of the significance attached to them. Conveying of ideas or information by means of drawings, maps, colorings, etc., then, is what is meant by graphic presentation.

There are many ways by which statistical data may be presented graphically. A few of the more suggestive of these will be illustrated.

#### BAR CHARTS

This chart, as is obvious, is the graphic presentation of data by bars. Two general types of bar charts are recognized, horizontal and vertical.

In Table XXIX there is shown the number of automobile fatalities in the city of Washington for a five-year period. It may be that in many such instances the only interest that the reader has in such statistics lies in the tendency toward a trend, if any, over the entire period. When this is true, some form of graphic presentation may well serve the purpose. The data in Table XXIX are presented as a horizontal bar chart in Chart 9,<sup>2</sup> and as a vertical bar chart in Chart 10.

<sup>1</sup> The demand, supply, and utility curves are being purposely omitted from this discussion because they are properly a part of economics.

<sup>2</sup> The particular arrangement of years on the vertical scale is more or less arbitrary as to sequence.

CHART 9. AUTOMOBILE FATALITIES IN THE CITY OF WASHINGTON,  
1922-1926

(Based on data in Table XXIX)

(HORIZONTAL BAR CHART)

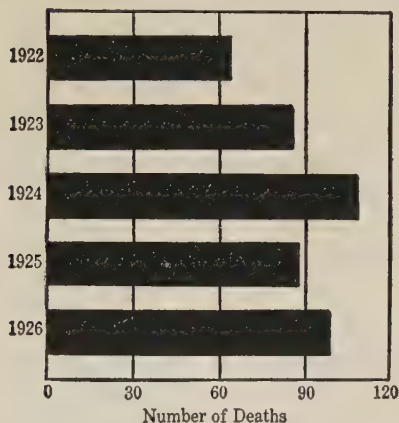


CHART 10. AUTOMOBILE FATALITIES IN THE CITY OF WASHINGTON,  
1922-1926

(Based on data in Table XXIX)

(VERTICAL BAR CHART)

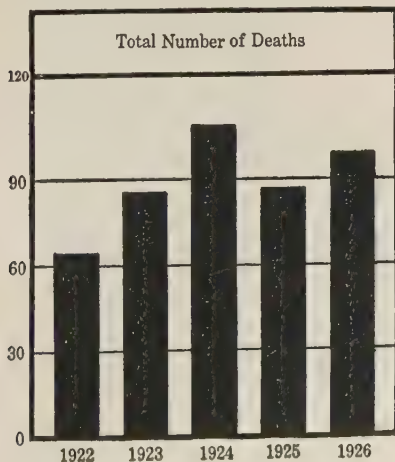


TABLE XXIX. AUTOMOBILE FATALITIES IN THE CITY OF WASHINGTON, 1922-1926 <sup>1</sup>

| YEAR  | TOTAL NUMBER OF DEATHS |
|-------|------------------------|
| 1922  | 64                     |
| 1923  | 86                     |
| 1924  | 108                    |
| 1925  | 88                     |
| 1926  | 98                     |
| Total | 444                    |

## MAPS

Statistical maps are of two general types. The first of these is the so-called "cross-hatch" map. This type is particularly useful when it is desired to present the data for a certain geographical area without regard to exact location within the area. For example, we may wish to show the relative importance of North and South Carolina, Georgia, and Florida in respect to cotton acreage harvested. In this instance no importance would be attached to the sections in those particular states where the cotton is grown. The object would be to show the relative position of each state as a whole.

TABLE XXX. ACREAGE OF HAY CROPS IN DELAWARE, BY COUNTIES, 1925 <sup>2</sup>

| COUNTY     | ACREAGE OF HAY CROPS |
|------------|----------------------|
|            | <i>Acres</i>         |
| Sussex     | 19,397               |
| Kent       | 25,705               |
| New Castle | 30,190               |
| Total      | 75,292               |

The data in Table XXX are chosen for illustrative purposes. Here is shown the acreage of hay crops in each

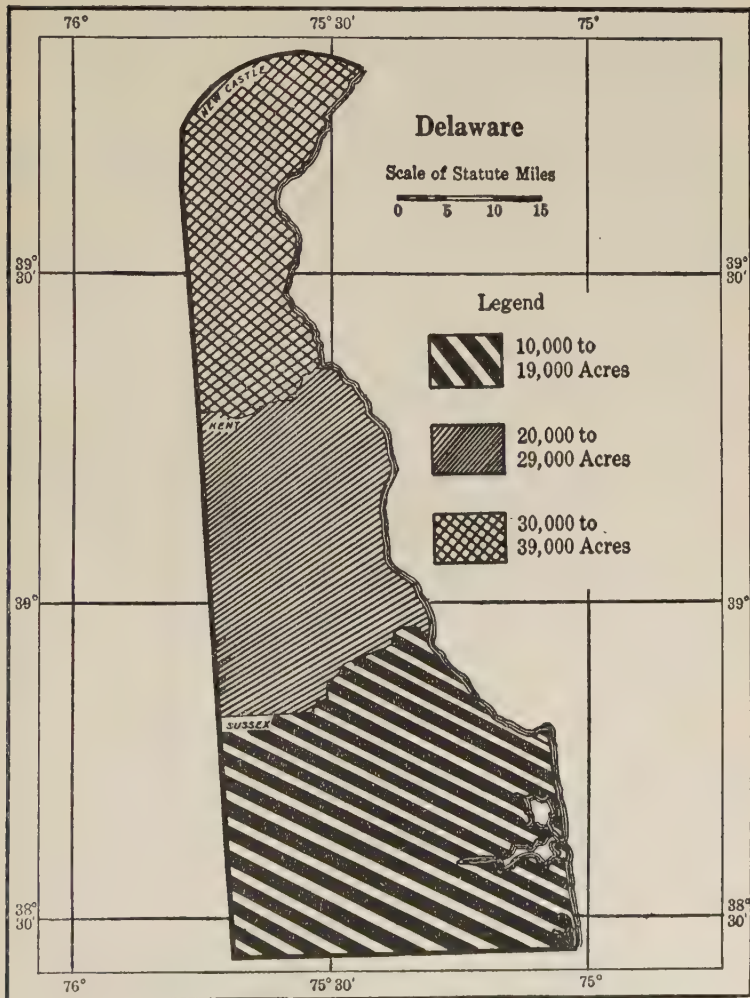
<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 379, Table 399.

<sup>2</sup> *United States Census of Agriculture, 1925, Part II*, page 85, Table IV.

CHART 11. ACREAGE OF HAY CROPS IN DELAWARE, BY COUNTIES, 1925

(Based on data in Table XXX)

(CROSS-HATCH MAP)



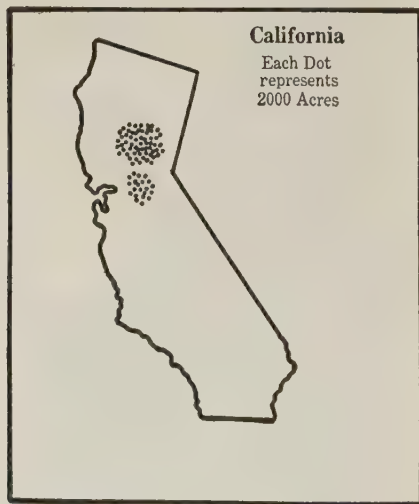
of the three counties of Delaware. In making a map to show the location of these acreages we are concerned only with the relative position that the counties bear to one another. We are not concerned with the distribution of the acreage of hay crops over the various sections within the several counties.

Chart 11 is a cross-hatch map showing by varied rulings and shadings the relative importance of each of Delaware's three counties in respect to acreage of hay crops in 1925.

CHART 12. APPROXIMATE LOCATION OF RICE ACREAGE IN CALIFORNIA IN 1927

(Based on original data in *Yearbook, United States Department of Agriculture, 1928*, page 747, Table 107)

(MULTIPLE DOT MAP)



The second of the two most commonly used types of maps is the multiple dot map. This is simply a map for a



certain geographical area that shows by several or more dots both the magnitude of the data and the exact or approximate location within the area. In Chart 12, for example, there is graphically shown both the acreage of rice in California for a specific period and its distribution or location in the particular sections of the state where it is grown. The latter constitutes the most desirable feature of the multiple dot map.

The single dot map has already been discussed in Chapter III under "Spatial Distribution."

### CIRCLE CHARTS

Statistical data are presented in a circle chart simply by dividing the circle into parts that are of the same relative size as the items to be presented. Let us assume that we have 100 apples, 50 of which are red, and the other 50 yellow. If we wish to present these data in a circle chart all that would be necessary would be to divide the circle into two equal parts. If there were 25 yellow apples, 25 green ones, and 50 red ones, then the circle would be divided into three parts. The yellow apples, constituting one-fourth of the total number of 100, would be represented by one-fourth of the area of the circle, and so on for the greens and reds.

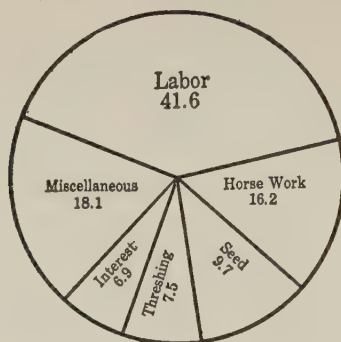
Chart 13 illustrates the dividing of a circle according to the relative magnitude of items included in it. Labor, for example, constitutes 41.6 per cent. of the total cost, and because of that the area of the circle showing the relative importance of this item constitutes 41.6 per cent. of the total area of the circle.

It is a common practice with some analysts to "cross-hatch" a circle chart in much the same way as illustrated by the map in Chart 11, page 69.

CHART 13. PERCENTAGE DISTRIBUTION OF GROSS FARM COSTS OF GROWING PEANUTS IN FLORIDA, 1925

(Based on original data contained in a report entitled *Peanuts, 1929, Report of the United States Tariff Commission to the President of the United States*, Appendix A, page 73, Table 59)

(OPEN-FACE CIRCLE CHART)



### HISTORIGRAM

In ordinary statistical analysis the common historigram is the most frequently used of all the many types of diagrams and charts in analyzing a historical or time series. This is not only because it is easily constructed, but also because it is a very satisfactory way of showing statistical facts that pertain to a definite period of time.

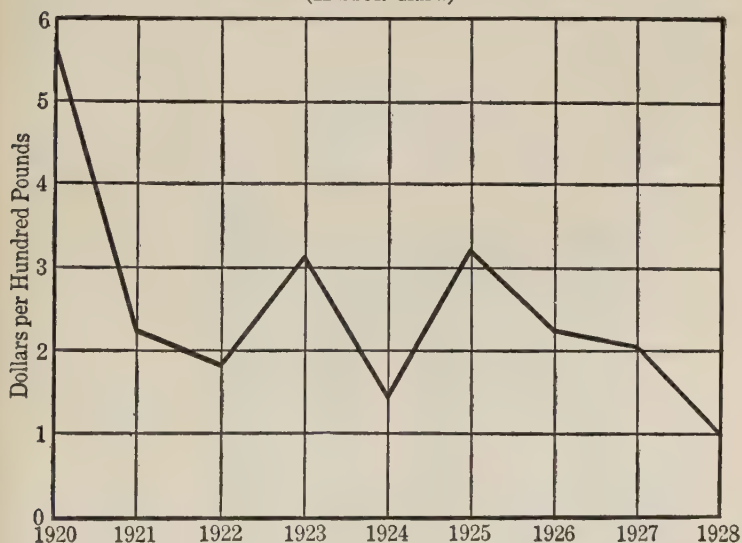
TABLE XXXI. AVERAGE L.C.L. PRICE OF POTATOES AT NEW YORK FOR THE MONTH OF JULY, 1920-1928<sup>1</sup>

| YEAR | PRICE PER HUNDRED POUNDS |
|------|--------------------------|
| 1920 | \$5.54                   |
| 1921 | 2.23                     |
| 1922 | 1.81                     |
| 1923 | 3.08                     |
| 1924 | 1.48                     |
| 1925 | 3.18                     |
| 1926 | 2.29                     |
| 1927 | 2.07                     |
| 1928 | 1.02                     |

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 816, Table 212.

Table XXXI shows the average price of potatoes at New York in July for a period of nine years, and in Chart 14 these prices are presented graphically by means of a line connecting the items for the individual years. This chart is a historigram, and it should not be confused with the histogram, in the construction of which the element of time is not a part.

CHART 14. AVERAGE L.C.L. PRICE OF POTATOES AT NEW YORK FOR  
THE MONTH OF JULY, 1920-1928  
(Based on data in Table XXXI)  
(HISTORIGRAM)



It will be observed that the reading of this chart is from two scales. The vertical scale, "Dollars per hundred pounds," is known as the y axis, while the horizontal scale, represented by the years from 1920 to 1928, is known as the x axis. These designations apply to any other kind of graph or chart similarly constructed.

Whenever it is desired to show the relation between two or more series of data it is a good plan to plot them, if the number is not too great, in one chart.<sup>1</sup> For example, in Chart 14 we could plot the receipts of potatoes at New York during the month of July for each respective year, using a dotted or colored line to distinguish it from the line representing price.

### CONSTITUENT-ELEMENT CHART

As the name suggests, a constituent-element chart is one which shows how or in what proportions the individual items of a total number are apportioned. In Table XXXII there are shown both separately and combined for five years the statistics of the acreage of tomatoes grown in the United States for the production of the fruit for consumption in the fresh state and for manufacture.

TABLE XXXII. ACREAGE OF TOMATOES HARVESTED IN THE UNITED STATES, 1925-1929 <sup>2</sup>

| YEAR | ACREAGE HARVESTED                                 |                             |                    |
|------|---------------------------------------------------|-----------------------------|--------------------|
|      | Tomatoes for<br>Consumption in<br>the Fresh State | Tomatoes for<br>Manufacture | Total              |
|      | <i>1,000 acres</i>                                | <i>1,000 acres</i>          | <i>1,000 acres</i> |
| 1925 | 134                                               | 350                         | 484                |
| 1926 | 111                                               | 262                         | 373                |
| 1927 | 139                                               | 256                         | 395                |
| 1928 | 145                                               | 254                         | 399                |
| 1929 | 149                                               | 285                         | 434                |

<sup>1</sup> When there is a lag in one series it is often advisable to plot each series separately so as to make it easier to determine the extent of this lag. This is readily done by placing the charts in a lighting box, or by merely holding the charts before a light and moving them back and forth until the position of greatest relationship is determined. It is understood, of course, that the coefficient of correlation with various lags must be determined before it can be definitely decided as to the proper lag.

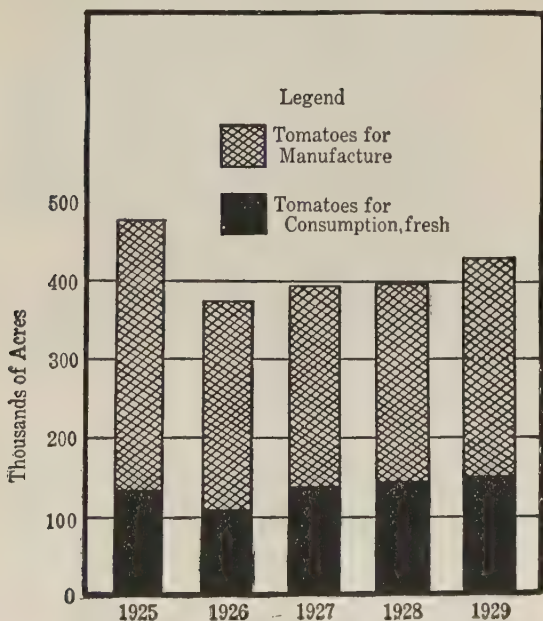
<sup>2</sup> *Yearbooks, United States Department of Agriculture; 1930, pages 781-2, Tables 276 and 277; 1928, page 821, Tables 223 and 224.*

Chart 15 is probably the best way of presenting these data in a constituent-element chart. The bars as constructed show precisely the apportionment of each of the two items. Chart 16 illustrates another way by which the same data may be presented.

CHART 15. ACREAGE OF TOMATOES HARVESTED IN THE UNITED STATES, 1925-1929

(Based on data in Table XXXII)

(CONSTITUENT-ELEMENT BAR CHART)



### SILHOUETTE CHART

The silhouette chart is useful when there are marked changes from one period to another, and when only one series of data is to be plotted in a single chart.

CHART 16. ACREAGE OF TOMATOES HARVESTED IN THE UNITED STATES,  
1925-1929

(Based on data in Table XXXII)  
(CONSTITUENT-ELEMENT BAND CHART)

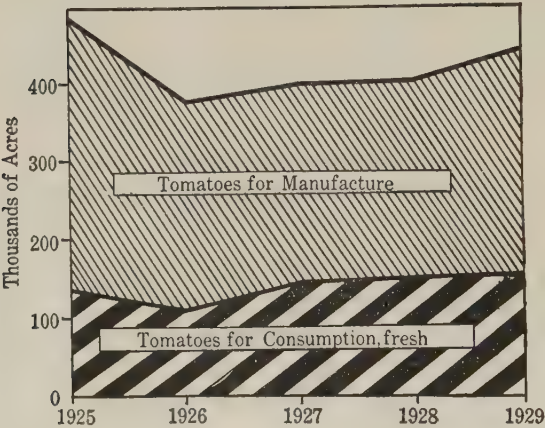


Table XXXIII shows the statistics of floor oilcloth imports for seven years. Chart 17 is a silhouette chart showing graphically the changes in the magnitude of these imports from year to year.

TABLE XXXIII. IMPORTS OF FLOOR OILCLOTH, 1922-1928 <sup>1</sup>

| YEAR | QUANTITY OF IMPORTS |
|------|---------------------|
|      | <i>Square Yards</i> |
| 1922 | 549                 |
| 1923 | 14,091              |
| 1924 | 8,259               |
| 1925 | 58,321              |
| 1926 | 27,053              |
| 1927 | 16,540              |
| 1928 | 7,605               |

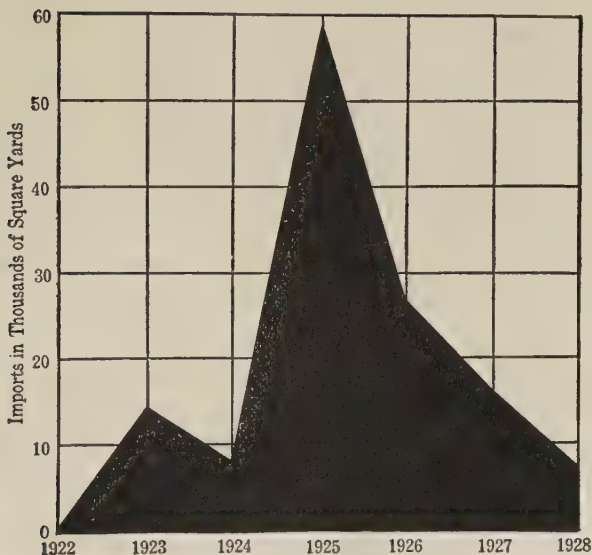
<sup>1</sup> *Summary of Tariff Information, 1929, Schedule 10, page 1659.*



CHART 17. IMPORTS OF FLOOR OILCLOTH, 1922-1928

(Based on data in Table XXXIII)

(SILHOUETTE CHART)



## EXCESS AND DEFICIT CHART

This type of chart is most commonly used in presenting such factors as net profit and loss, net imports and exports, and the like. As will be observed, the excess and deficit chart cannot be used if it is desired to show actual or absolute quantities. This is because it is constructed from net quantities, or the difference between items.

Table XXXIV will make this clear. There are shown in this table the statistics of actual quantities of buckwheat imported into and exported from the United States. The statistics of net imports and net exports are also shown, and it is these data that are plotted in Chart 18.

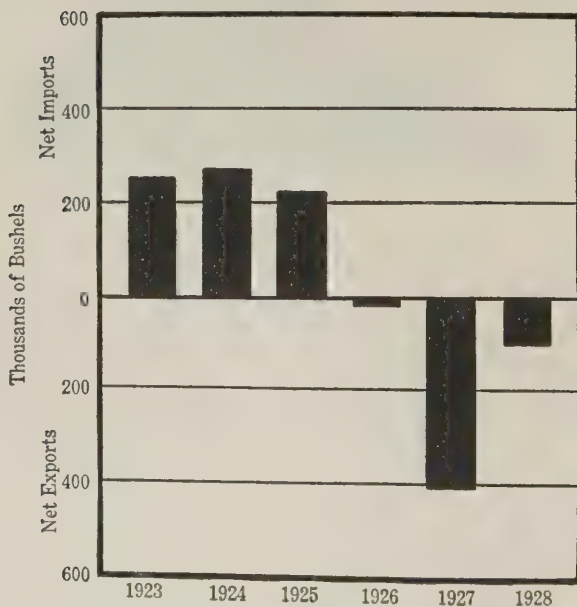
TABLE XXXIV. IMPORTS AND EXPORTS OF BUCKWHEAT, 1923-1928 <sup>1</sup>

| YEAR | IMPORTS INTO<br>THE UNITED<br>STATES | EXPORTS FROM<br>THE UNITED<br>STATES | NET IMPORTS<br>(Excess of<br>Imports over<br>Exports) | NET EXPORTS<br>(Excess of<br>Exports over<br>Imports) |
|------|--------------------------------------|--------------------------------------|-------------------------------------------------------|-------------------------------------------------------|
|      | <i>Bushels</i>                       | <i>Bushels</i>                       | <i>Bushels</i>                                        | <i>Bushels</i>                                        |
| 1923 | 293,000                              | 49,837                               | 243,163                                               | —                                                     |
| 1924 | 385,000                              | 121,696                              | 263,304                                               | —                                                     |
| 1925 | 336,000                              | 109,694                              | 226,306                                               | —                                                     |
| 1926 | 89,000                               | 101,389                              | —                                                     | 12,389                                                |
| 1927 | 63,000                               | 466,500                              | —                                                     | 403,500                                               |
| 1928 | 66,000                               | 160,545                              | —                                                     | 94,545                                                |

CHART 18. IMPORTS AND EXPORTS OF BUCKWHEAT, 1923-1928

(Based on data in Table XXXIV)

(EXCESS AND DEFICIT CHART)

<sup>1</sup> Summary of Tariff Information, 1929, Schedule 7, page 1181.

## RATIO CHART

When it is desired to show not only the actual size of items, but also the ratio between individuals and each succeeding change, the ordinary equal-spaced paper will not serve the purpose. This is because the equal-spaced paper is ruled to show actual changes from period to period, and not the ratio that these changes bear to preceding items. Ratio paper, on the other hand, is ruled in such a way as to show both the actual magnitude of each item and to indicate the proportionate change in each item for the period. Furthermore, the ruling of the paper is such that the same percentage changes are always represented by equal vertical distances.

In Table XXXV are data purporting to show the earnings of two groups of individuals over the same period. The earnings of the first group, it will be seen, doubled each month, representing an increase of 100 per cent. for each month over the immediately preceding month. These earnings are shown graphically by the continuous

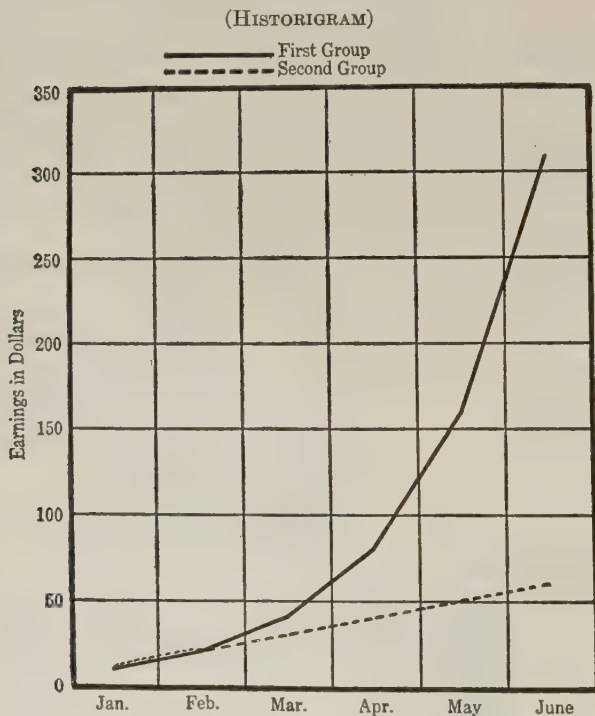
TABLE XXXV. EARNINGS OF TWO GROUPS OF INDIVIDUALS FOR SIX MONTHS <sup>1</sup>

| MONTH    | FIRST GROUP        |                               |                                         | SECOND GROUP       |                               |                                         |
|----------|--------------------|-------------------------------|-----------------------------------------|--------------------|-------------------------------|-----------------------------------------|
|          | Amount of Earnings | Increase over Preceding Month | Per Cent. Increase over Preceding Month | Amount of Earnings | Increase over Preceding Month | Per Cent. Increase over Preceding Month |
| January  | \$10               | —                             | —                                       | \$10               | —                             | —                                       |
| February | 20                 | \$10                          | 100                                     | 20                 | \$10                          | 100                                     |
| March    | 40                 | 20                            | 100                                     | 30                 | 10                            | 50                                      |
| April    | 80                 | 40                            | 100                                     | 40                 | 10                            | 33                                      |
| May      | 160                | 80                            | 100                                     | 50                 | 10                            | 25                                      |
| June     | 320                | 160                           | 100                                     | 60                 | 10                            | 20                                      |

<sup>1</sup> Assumed data.

line in Charts 19 and 20 respectively. The reader will observe at a glance that it is Chart 20, the ratio chart, which shows both the actual earnings and the constant percentage change from month to month.

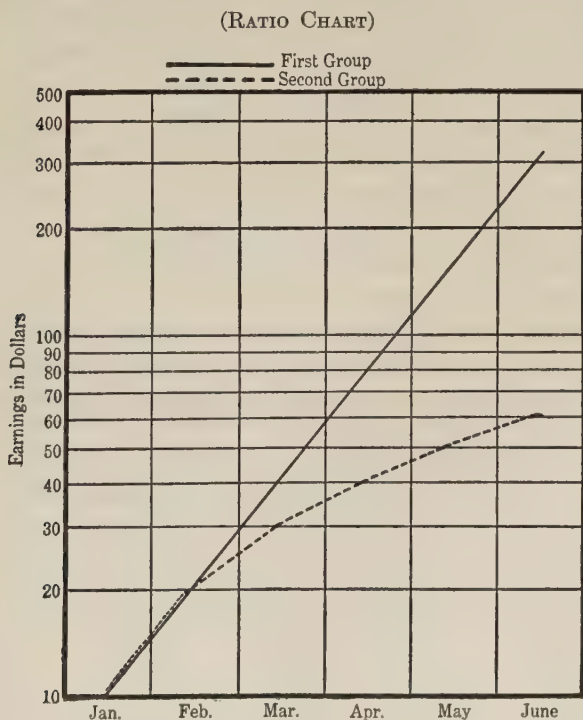
CHART 19. EARNINGS OF TWO GROUPS OF INDIVIDUALS FOR SIX MONTHS  
(Based on data in Table XXXV)



The earnings of the second group increased 10 dollars each month, but the percentage increase showed a decline throughout the period from 100 per cent. in February to 20 per cent. in June. In Chart 19 the earnings are represented by the straight dotted line, which does not furnish

CHART 20. EARNINGS OF TWO GROUPS OF INDIVIDUALS FOR SIX MONTHS

(Based on data in Table XXXV. The horizontal lines extending parallel to the x axis are drawn at distances apart in proportion to the size of the logarithms of the numbers on the y axis. This is the principle underlying all such logarithmic scales. In this chart the rulings have been made only for the y axis. The paper, then, is known as semi-logarithmic. When both horizontal and vertical lines are thus drawn it is called logarithmic paper.)



any index as to the percentage change from month to month. Chart 20, however, shows by the dotted line not only the actual earnings, but also indicates the percentage or "ratio" change in these earnings from month to month. It is this type of chart, therefore, that is more satisfactory for presenting ratios or percentage changes.

## SCATTER DIAGRAM

A scatter diagram is merely a chart showing paired relationships. That is to say, it is so constructed that each value on the x axis is plotted at the vertical point where a y value intersects. A study of Tables XXXVI and XXXVII and of Chart 21 will make this more easily understood. In Table XXXVI are the data on precipitation at Norfolk, Virginia, for the three summer months during the years from 1920 to 1928. Table XXXVII shows the statistics of yield of peanuts per acre in Virginia for this same period of years. The problem is to plot these paired values in a scatter diagram. Chart 21 illustrates how this is done. Precipitation, it will be seen, represents the x values, and yield per acre the y values. Looking at Table XXXVI we find that in 1922, for example, the average precipitation for the three summer months was 9.91 inches, and in Table XXXVII we see

TABLE XXXVI. PRECIPITATION AT NORFOLK, VIRGINIA, FOR THREE MONTHS, 1920-1928 <sup>1</sup>

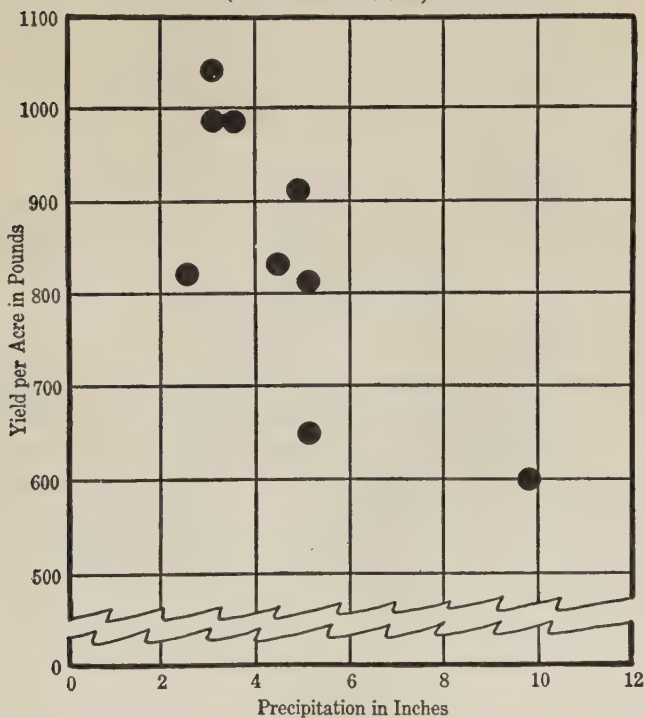
| YEAR | PRECIPITATION |               |               |                          |
|------|---------------|---------------|---------------|--------------------------|
|      | June          | July          | August        | Average for Three Months |
|      | <i>Inches</i> | <i>Inches</i> | <i>Inches</i> | <i>Inches</i>            |
| 1920 | 5.05          | 4.33          | 3.83          | 4.40                     |
| 1921 | 1.05          | 3.27          | 3.13          | 2.48                     |
| 1922 | 9.78          | 11.92         | 8.04          | 9.91                     |
| 1923 | 1.43          | 4.14          | 4.47          | 3.35                     |
| 1924 | 6.46          | 4.45          | 4.27          | 5.06                     |
| 1925 | 2.60          | 5.52          | 1.05          | 3.06                     |
| 1926 | 2.69          | 1.03          | 5.73          | 3.15                     |
| 1927 | 4.68          | 6.72          | 3.85          | 5.08                     |
| 1928 | 7.84          | 4.41          | 2.05          | 4.77                     |
| Mean | —             | —             | —             | 4.58                     |

<sup>1</sup> Yearbooks, United States Department of Agriculture; 1928, page 1116, Table 592; 1927, pages 1209-11, Table 557; 1926, pages 1270-71, Table 568; 1925, pages 1504-06, Table 796.



CHART 21. SCATTER DIAGRAM OF PRECIPITATION FOR THREE MONTHS AT NORFOLK, VIRGINIA, AND YIELD OF PEANUTS PER ACRE IN VIRGINIA, 1920-1928

(Based on data in Tables XXXVI and XXXVII)  
(SCATTER DIAGRAM)



that the yield of peanuts per acre in 1922 was 600 pounds. The precipitation and yield per acre, then, are the paired values.

In plotting the intersection of these values in the scatter diagram we simply locate the value of 9.91 on the x axis and move upward on the y axis until the value of 600 is reached. The other intersections of paired values are located in the same manner.

TABLE XXXVII. YIELD OF PEANUTS PER ACRE IN VIRGINIA, 1920-1928 <sup>1</sup>

| YEAR | YIELD PER ACRE |
|------|----------------|
|      | <i>Pounds</i>  |
| 1920 | 830            |
| 1921 | 820            |
| 1922 | 600            |
| 1923 | 990            |
| 1924 | 650            |
| 1925 | 1,040          |
| 1926 | 990            |
| 1927 | 810            |
| 1928 | 910            |
| Mean | 849            |

## REGRESSION CHART

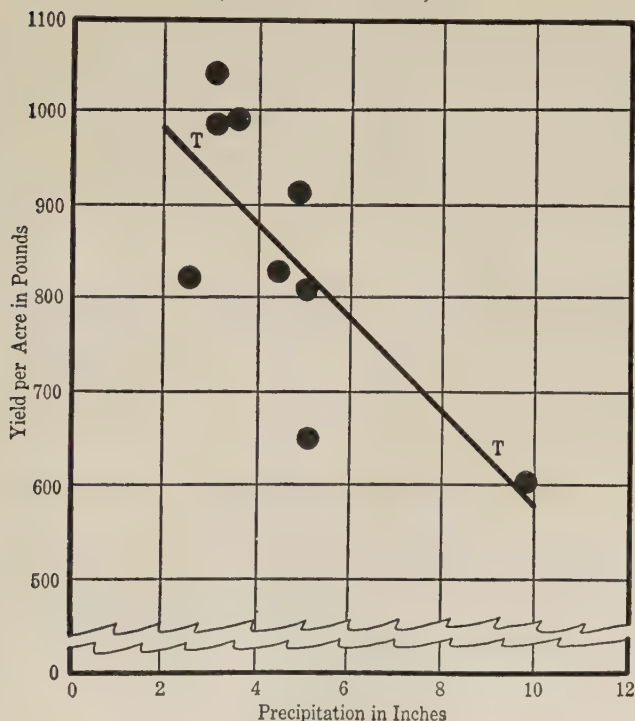
The regression chart is constructed to show the closest relationship between corresponding items in two or more series, with this relationship being expressed by either a straight line or curve. Since the ordinates of such a line are based on rigid mathematical calculations involving statistical terms that have not yet been defined in this text, a more detailed explanation of fitting the line will be reserved for a later chapter. The principal reason for considering the regression chart at this time is to enable the student to gain some conception of its nature along with the other common forms of graphic presentation. Chart 22, therefore, showing a straight line relationship, and based on data in Tables XXXVI and XXXVII, is designed to serve this purpose. As will be observed, the y axis represents yield of peanuts per acre, while the x axis represents precipitation, just as in Chart 21. The characteristic feature of Chart 22 is that the straight

<sup>1</sup> Yearbooks, United States Department of Agriculture; 1928, page 871, Table 302; 1927, page 947, Table 292; 1924, page 794, Table 370; 1923, page 839, Table 351; 1922, page 759, Table 310; 1921, page 645, Table 243.

CHART 22. RELATION BETWEEN YIELD OF PEANUTS IN VIRGINIA AND PRECIPITATION FOR THREE MONTHS AT NORFOLK, 1920-1928

(Based on data in Tables XXXVI and XXXVII)

(REGRESSION CHART)



line TT shows the highest average relationship between yield and precipitation that can be represented by a straight line constructed or fitted in this manner. It shows, for example, that with a precipitation of 10 inches the yield on an average is likely to be about 580 pounds per acre. Any other yield that is most likely to be associated with a given precipitation is determined by locating the y value of the regression line at the intersection of the

x and y axis. Accordingly, if we wish to estimate the yield that is most likely to be associated with a precipitation of 6 inches we merely move upward on the y axis from 6 on the x axis to the point of intersection with the line. The value on the y axis that corresponds to this point of intersection is the most probable yield, or the most probable y value that is likely to be associated with the given x value. The mathematics involved in calculating the ordinates of the regression line is presented in detail in Chapter XIII, pages 248 to 261.

Another excellent way of presenting relationships between series is to plot the multiples of standard deviation. These multiples may be calculated from the actual values, first differences, deviations of actual values from their mean, or deviations of percentage changes from their mean. In addition to these, the actual values of each series may be plotted in ordinary line charts.

### FREQUENCY GRAPHS

Three common types of graphs are used in presenting frequency distributions, the histogram, the polygon, and the smoothed curve. A brief explanation will show the distinction between these graphs.

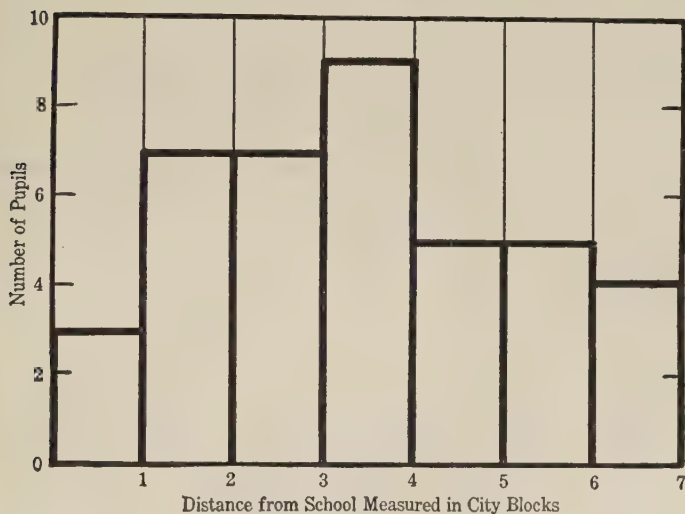
#### *Histogram*

The histogram is often referred to as a column diagram. It is merely a column or series of columns designed to show the relative magnitudes of the y values that are associated with varying x values. The time element does not enter into the construction of a histogram, and it is in this respect that it may be distinguished from the histogram. Let us illustrate its construction by plotting the data in Table XXI, Chapter III, page 55.

CHART 23. NUMBER OF PUPILS LIVING WITHIN SPECIFIED DISTANCES OF SCHOOL

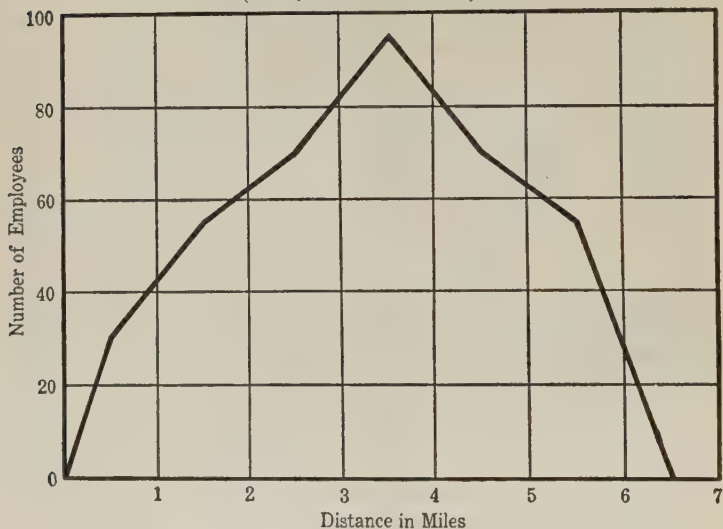
(Based on data in Table XXI)

(HISTOGRAM)

*Frequency Polygon*

In plotting the frequency polygon we merely draw a continuous line connecting the tops of the midpoints of columns in the histogram. The method of constructing a polygon is illustrated in Chart 24. As will be observed, the ordinates on the y axis are plotted midway between the values on the x axis, the assumption being that the frequencies are evenly distributed about the midpoints of the x values. The polygon as constructed in Chart 24 shows that no employee lives as far as  $6\frac{1}{2}$  miles from his office. This type of chart may sometimes be useful when there is a large number of frequencies in the classes, and even when the frequencies are few it may serve a purpose.

CHART 24. NUMBER OF FEDERAL EMPLOYEES LIVING WITHIN SPECIFIED  
DISTANCES OF THEIR OFFICES  
(Based on assumed data)  
(FREQUENCY POLYGON)



*Smoothed Frequency Curve*

The smoothed curve is drawn in such a way that the area it includes is approximately the same as the area in the frequency polygon. If we were smoothing the polygon in Chart 24, for instance, the curve should be so superimposed upon the polygon that the area under the line is about equal to the area covered by the polygon itself. It may be stated also that whenever possible the area of the individual columns in the histogram should not be changed by the smoothing process. This, however, cannot always be observed, as exemplified by Chart 23. Smoothing of this histogram would necessarily change the areas of some of the columns because of the nature of the distribution.



DIMINISHING RETURNS <sup>1</sup>

By the law of diminishing returns we mean that additional units of expenditure yield a less than proportionate increase in product. Let us assume, for example, that a corn grower in Connecticut has 8 plots of corn land of equal size, which, for our purposes, we will consider to be of equal fertility, and equal in all other aspects that are related to yield. During a certain season we will say the following applications of fertilizer were made, with resulting yields of grain as indicated:

| PLOT NUMBER* | APPLICATION<br>OF<br>FERTILIZER | YIELD OF GRAIN<br>PER PLOT | INCREASE IN<br>YIELD OF<br>ONE PLOT<br>OVER THE<br>OTHER | AVERAGE YIELD<br>PER 100 POUNDS<br>OF FERTILIZER<br>APPLIED |
|--------------|---------------------------------|----------------------------|----------------------------------------------------------|-------------------------------------------------------------|
|              | <i>Pounds</i>                   | <i>Bushels</i>             | <i>Bushels</i>                                           | <i>Bushels</i>                                              |
| 1            | 0                               | 0                          | 0                                                        | 0                                                           |
| 2            | 100                             | 5                          | 5                                                        | 5.00                                                        |
| 3            | 200                             | 15                         | 10                                                       | 7.50                                                        |
| 4            | 300                             | 27                         | 12                                                       | 9.00                                                        |
| 5            | 400                             | 38                         | 11                                                       | 9.50                                                        |
| 6            | 500                             | 48                         | 10                                                       | 9.60                                                        |
| 7            | 600                             | 58                         | 10                                                       | 9.67                                                        |
| 8            | 700                             | 60                         | 2                                                        | 8.57                                                        |

\* Assumed data.

It is seen that when no fertilizer was applied to plot 1 there was no production of grain. An application of 100 pounds of fertilizer was followed by a yield on plot 2 of 5 bushels. Plot 3, which received an application of 200 pounds of fertilizer, yielded 15 bushels of grain, an increase of 10 bushels over plot 2, and so on for the remaining

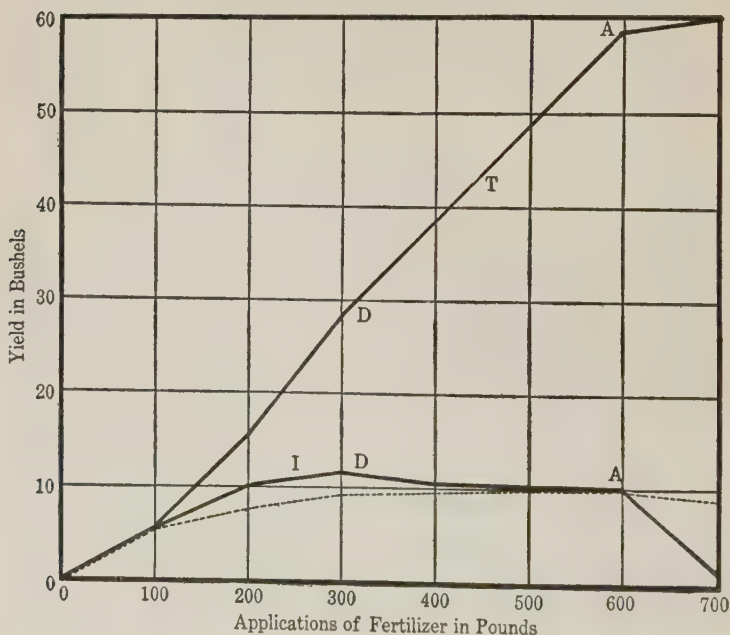
<sup>1</sup> There are many kinds of curves that may be drawn to represent diminishing returns. Some of these are very complicated, and have no place in an elementary text, since they involve so many ramifications. This applies also to curves representing increasing returns. Only the simplest illustrations, therefore, are used herein. It is to be understood by the student that in speaking of increasing, decreasing, constant, and highest average returns it is ordinarily the quantity of returns that is referred to, and not the value, unless, of course, otherwise specified.

plots. We see that the yield on plot 4 was 12 bushels greater than that of plot 3. In the case of plot 5, however, the yield was only 11 bushels greater than that of plot 4. It is at this point, therefore, that the yield per application of 100 pounds of fertilizer was proportionately less than the yield obtained from a smaller application. In other words, an increased application of 100 pounds on plot 4 gave an increased yield of 12 bushels, while on plot 5 the increased application of 100 pounds was associated with an increase in yield of only 11 bushels. With the application of the

CHART 25. DIMINISHING RETURNS FROM INCREASING APPLICATIONS OF FERTILIZER

(Based on assumed data)

(DIMINISHING RETURNS)



400 pounds of fertilizer, then, the yield was proportionately less than the yield from plot 4, which received 300 pounds of fertilizer. It will be seen, therefore, that proceeding from the 300-pound application there was a proportionate decrease in yield.

The preceding chart shows more clearly the nature of diminishing or decreasing returns. The continuous line designated as T represents the absolute total yield, while the continuous line designated as I represents the increase in yield from each successive increase in fertilizer application. The point of diminishing returns on lines T and I is designated as D.

#### HIGHEST AVERAGE RETURNS

The point of diminishing returns must not be construed to mean that it coincides always with the highest average returns. Returns may diminish and yet the average returns may increase for a time. This is because, even though a successive expenditure may result in an increased return that is not as great as the increased return from the immediately preceding expenditure, the average returns may show an increase. Let us turn back to the data from which Chart 25 was constructed. Reading down the column in which is recorded the average yield per 100 pounds of fertilizer applied, we see that the highest average was obtained from the plot on which 600 pounds of fertilizer were applied. This average is 9.67, meaning that for each 100 pounds of fertilizer applied there was an average return of 9.67 bushels. No other application of fertilizer gave an average yield as great as that associated with the application of 600 pounds made to plot 7. In Chart 25 the average returns are represented by the dotted line. The point of highest average returns on the continuous

line T and the dotted line is the ordinate of the y axis corresponding to the x value of 600. This point is designated as A on both the dotted line and on Curve T.

In connection with the highest average returns it is important to remember that when the total returns are plotted as a smoothed curve and the point of diminishing returns located, the area below that part of the curve is the same as the area below the smoothed curve of average returns up to the point of maximum. For instance, if we draw a perpendicular straight line from the point of diminishing returns as located on the smoothed curve of total returns to the x axis and measure that part of the area under the curve lying to the left of this perpendicular line, we will find it to be the same as that part of the area under the curve representing average returns that lies to the left of a straight perpendicular line drawn from the point of highest average to the x axis.

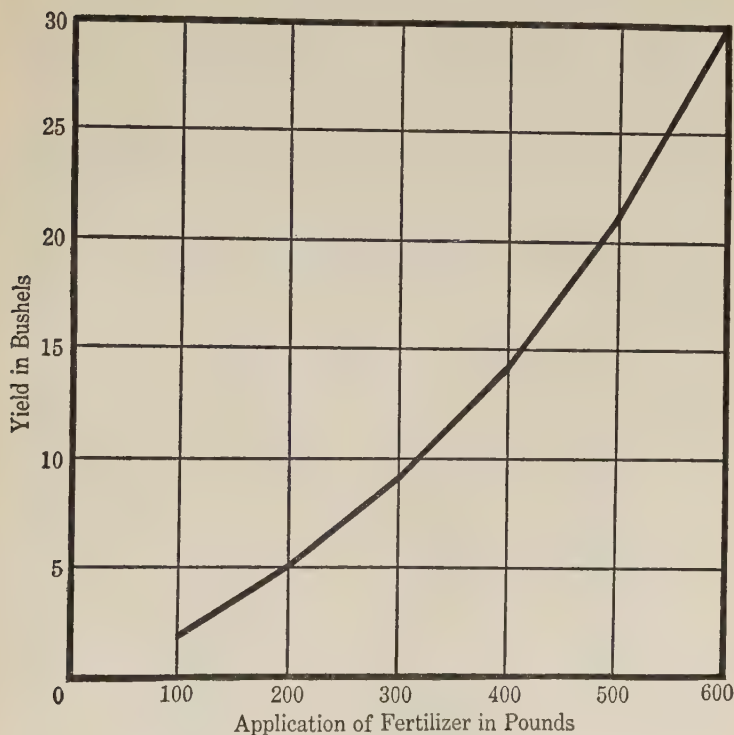
### INCREASING RETURNS

Whenever an additional unit of expenditure is followed by a return proportionately greater than that resulting from the immediately preceding unit of expenditure, we have what is commonly known as increasing returns. For example, if a Texas farmer in the trucking area finds that for each additional \$10.00 he expends per acre on an onion crop he increases his yield by 25, 40, 60, and 90 bushels, his returns are said to be increasing because every time he expends an additional \$10.00 his returns are more than proportionately increased. The chart on page 93 indicates the general nature of increasing returns. It shows that the yield resulting from each increased application of fertilizer was proportionately greater than the increases from the smaller applications.

CHART 26. INCREASING YIELDS OF WHEAT RESULTING FROM SUCCESSIVE APPLICATIONS OF FERTILIZER

(Based on assumed data)

(INCREASING RETURNS)

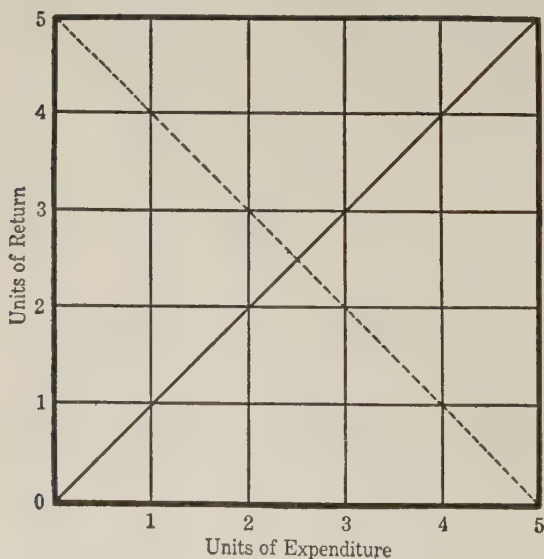
CONSTANT RETURNS <sup>1</sup>

If the returns increase or decrease the same amount with each successive unit of expenditure, so that the amount of

<sup>1</sup> The line of fixed returns might also be illustrated, but since fixed returns from increased expenditures are usually the result of a lack of proper proportionment of productive factors it is being purposely omitted. It will be readily understood that if a manufacturer increases his factory personnel by 10 every week without increasing his output there is a lack of proper adjustment in the number of machines, materials, or some other phase of production.

increase or decrease is equally proportionate in each case to expenditure, they are said to be constant. Chart 27 illustrates both increasing and decreasing constant returns, the former being represented by the continuous line extending from the lower left of the chart to the upper right, and the latter by the dotted line extending from the upper left of the chart to the lower right.

CHART 27.    CONSTANT INCREASING AND DECREASING RETURNS  
                   (Based on assumed data)  
                   (CONSTANT RETURNS)





## CHAPTER V

### THE ARITHMETIC MEAN

The arithmetic mean is the most commonly used of all the averages, and it is probably more readily understood than any other. This is due largely, perhaps, to the ease of calculation and to the fact that the apparent significance of this particular average is easily appreciated. It is undoubtedly true that the ease with which the arithmetic mean is understood can be attributed to the readiness with which it is calculated, since the latter factor is largely responsible for its wide application in the fields of education, biology, and economics.

TABLE XXXVIII. PRODUCTION OF PEANUTS IN THE UNITED STATES,  
1917-1926 <sup>1</sup>

| YEAR  | PRODUCTION          | DEVIATION FROM THE MEAN |                     |
|-------|---------------------|-------------------------|---------------------|
|       |                     | Minus                   | Plus                |
|       | <i>1,000 pounds</i> | <i>1,000 pounds</i>     | <i>1,000 pounds</i> |
| 1917  | 1,432,581           | —                       | 584,283.8           |
| 1918  | 1,240,102           | —                       | 391,804.8           |
| 1919  | 783,273             | 65,024.2                | —                   |
| 1920  | 841,474             | 6,823.2                 | —                   |
| 1921  | 829,307             | 18,990.2                | —                   |
| 1922  | 633,114             | 215,183.2               | —                   |
| 1923  | 647,762             | 200,535.2               | —                   |
| 1924  | 745,059             | 103,238.2               | —                   |
| 1925  | 698,475             | 149,822.2               | —                   |
| 1926  | 631,825             | 216,472.2               | —                   |
| Mean  | 848,297.2           | —                       | —                   |
| Total | —                   | 976,088.6               | 976,088.6           |

<sup>1</sup> Yearbooks, United States Department of Agriculture; 1927, page 947, Table 292; 1926, page 997, Table 297.

The arithmetic mean may be defined as the point of central tendency from which the algebraic sum of the deviations is zero. Therefore, if the deviations of individual items from their mean are summated algebraically, the plus deviations summated in one column and the minus deviations in another, the summations will be of the same size, with a difference of zero, showing that the mean is the point around which the deviations reach a minimum. This principle is illustrated in the preceding table.

The average production of peanuts for the ten-year period, 1917-1926, inclusive, was 848,297.2 thousands of pounds. As will be seen in the preceding table, the algebraic sum of the deviations from the mean is zero, since the totals of the minus and plus deviations are of the same magnitude. This is the equivalent of saying that the total difference between the mean and items smaller than the mean is the same as the total difference between the mean and items larger than the mean.

TABLE XXXIX. TEACHERS IN VOCATIONAL SCHOOLS IN THE UNITED STATES, 1924-1927 <sup>1</sup>

| YEAR  | TEACHERS IN VOCATIONAL SCHOOLS |
|-------|--------------------------------|
|       | <i>Number</i>                  |
| 1924  | 16,192                         |
| 1925  | 17,546                         |
| 1926  | 18,717 *                       |
| 1927  | 18,900 *                       |
| Total | 71,355                         |
| Mean  | 17,839                         |

\* Includes Hawaii.

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 120, Table 128. Data relate only to institutions federally aided under the act commonly known as the Smith-Hughes Act, or the National Vocational Educational Act.

## SIMPLE ARITHMETIC MEAN

The simple arithmetic mean is merely the result obtained when the sum of a series or array of items is divided by the number of items. Table XXXIX illustrates its calculation.

In the preceding calculation the mean is obtained by merely summing the items and dividing by the number, which is four in this case. The formula for this calculation

is  $\frac{\Sigma X}{N}$ , in which  $\Sigma X$  is the sum of the items and  $N$  the

number of items. Another formula,  $\frac{1}{N} \Sigma X$ , may be used in

calculating the mean. In this formula,  $\frac{1}{N}$  is the fractional

part of the number of items and  $\Sigma X$  is the sum of the values or items in the series. Calculation of the mean from this formula is done by simply dividing the number of items into 1 and then multiplying by the sum of the items. In Table XXXIX there are 4 items, and when 4 is divided into 1 a value of .25 is obtained. This value when multiplied by 71,355, the sum of the items, gives 17,839, the mean.

It is well to remember that although the simple arithmetic mean shows the average size of items in a series or array it furnishes no indication as to the extent of variation in magnitudes. Table XL, for example, shows two series with means of identical size, but with considerable variation in the size of corresponding items. In the first series the items range from 2.92 to 4.82, while in the other the range is from 0.62 to 6.90. This indicates that there are considerable variations in the magnitudes of items associated with the same periods of time, and this is clearly shown in the following table:

TABLE XL. AVERAGE PRECIPITATION AT SPECIFIED PLACES IN MARYLAND AND OREGON FOR TWELVE MONTHS <sup>1</sup>

| MONTH          | PRECIPITATION AT<br>BALTIMORE, MARYLAND | PRECIPITATION AT<br>PORTLAND, OREGON |
|----------------|-----------------------------------------|--------------------------------------|
|                | <i>Inches</i>                           | <i>Inches</i>                        |
| January        | 3.22                                    | 6.59                                 |
| February       | 3.51                                    | 5.42                                 |
| March          | 3.88                                    | 4.66                                 |
| April          | 3.27                                    | 3.02                                 |
| May            | 3.56                                    | 2.23                                 |
| June           | 3.84                                    | 1.64                                 |
| July           | 4.82                                    | 0.62                                 |
| August         | 4.21                                    | 0.63                                 |
| September      | 3.85                                    | 1.84                                 |
| October        | 3.02                                    | 3.28                                 |
| November       | 2.92                                    | 6.41                                 |
| December       | 3.08                                    | 6.90                                 |
| Annual average | 3.60                                    | 3.60                                 |

The simple arithmetic mean may be calculated from an assumed mean. In this case a number is selected, preferably one that lies within the range of the items, and the

TABLE XLI. UNIVERSITIES, COLLEGES, AND PROFESSIONAL SCHOOLS IN CONTINENTAL UNITED STATES, 1890-1926 <sup>2</sup>

| YEAR         | NUMBER OF<br>INSTITUTIONS | DEVIATION FROM ASSUMED MEAN |      |
|--------------|---------------------------|-----------------------------|------|
|              |                           | Minus                       | Plus |
| 1890         | 657                       | 43                          | —    |
| 1900         | 664                       | 36                          | —    |
| 1910         | 602                       | 98                          | —    |
| 1920         | 670                       | 30                          | —    |
| 1922         | 780                       | —                           | 80   |
| 1924         | 913                       | —                           | 213  |
| 1926         | 975                       | —                           | 275  |
| Total        | 5,261                     | 207                         | 568  |
| Assumed mean | 700                       | —                           | —    |
| True mean    | 752                       | —                           | —    |

<sup>1</sup> *Statistical Abstract of the United States, 1928*, pages 139 and 143, Table 146. Figures are long-time averages.

<sup>2</sup> *Statistical Abstract of the United States, 1928*, page 114, Table 122.

deviations of individual items are determined, the plus and minus deviations being summated in separate columns. The difference between the sums of these values is then divided by the number of items, and the product thus obtained, which is the correction factor, is added to or subtracted from the assumed mean, as the case may be, to obtain the true mean of the series. Table XLI illustrates the calculation of the mean by this method.

The sum of the minus deviations is 207, while the sum of the plus deviations is 568. Since these summations are of different magnitudes it is immediately obvious that the assumed mean does not coincide with the true mean, for we have already seen that the sums of the minus and plus deviations from the true mean are such that the difference between them is zero. It becomes necessary, therefore, to make a correction for the true mean in Table XLI. This is a very simple matter. Subtracting 207, the sum of the minus deviations, from 568, the sum of the plus deviations, leaves a plus quantity of 361, which divided by 7, the number of items, gives 52 as the correction factor. This being a plus quantity, it is added to 700, the assumed mean, to obtain the true mean of 752.

### WEIGHTED MEAN

There are cases when it is desired to give full weight to variations. We find, for example, by referring to Tables XLII and XLIII, that in 1927 the yield of corn per acre in the North Central States ranged from 25.0 bushels in North Dakota to 35.5 bushels in Iowa, and that the acreage of corn ranged from 959,000 acres in North Dakota to 10,901,000 acres in Iowa. Obviously, there are

two ways of calculating an average yield. We may simply summate the yields in the several states and take the average of that sum. This, of course, would merely be the simple average. Table XLII illustrates how this is done.

TABLE XLII. YIELD OF CORN PER ACRE IN THE NORTH CENTRAL STATES, 1927 <sup>1</sup>

| STATE                            | YIELD PER ACRE |
|----------------------------------|----------------|
|                                  | <i>Bushels</i> |
| North Dakota                     | 25.0           |
| Michigan                         | 27.5           |
| South Dakota                     | 29.0           |
| Missouri                         | 29.0           |
| Illinois                         | 30.0           |
| Kansas                           | 30.0           |
| Minnesota                        | 30.5           |
| Indiana                          | 31.5           |
| Ohio                             | 32.5           |
| Wisconsin                        | 32.5           |
| Nebraska                         | 33.1           |
| Iowa                             | 35.5           |
| Simple average yield<br>per acre | 30.5           |

The other way of calculating the average yield per acre is to summate both acreage and production in the several states, and then divide the total production by the total acreage. The result thus obtained is the weighted average yield per acre. Table XLIII shows the statistics of acreage and production for the individual states and for the group of states as a whole. Dividing the total production by the total acreage the weighted average yield per acre is found to be 31.5 bushels, as compared with an unweighted average yield of 30.5 bushels in Table XLII.

<sup>1</sup> Calculated from data in Table XLIII, page 101.



## THE ARITHMETIC MEAN

101

TABLE XLIII. ACREAGE AND PRODUCTION OF CORN IN THE NORTH CENTRAL STATES, 1927 <sup>1</sup>

| STATE            | ACREAGE<br>HARVESTED | TOTAL<br>PRODUCTION  | YIELD PER<br>ACRE |
|------------------|----------------------|----------------------|-------------------|
|                  | <i>1,000 Acres</i>   | <i>1,000 Bushels</i> | <i>Bushels</i>    |
| North Dakota     | 959                  | 23,975               | 25.0              |
| Michigan         | 1,418                | 38,995               | 27.5              |
| South Dakota     | 4,655                | 134,995              | 29.0              |
| Missouri         | 5,796                | 168,084              | 29.0              |
| Illinois         | 8,469                | 254,070              | 30.0              |
| Kansas           | 5,897                | 176,910              | 30.0              |
| Minnesota        | 4,172                | 127,246              | 30.5              |
| Indiana          | 4,205                | 132,458              | 31.5              |
| Ohio             | 3,376                | 109,720              | 32.5              |
| Wisconsin        | 2,100                | 68,250               | 32.5              |
| Nebraska         | 8,805                | 291,446              | 33.1              |
| Iowa             | 10,901               | 386,986              | 35.5              |
| Total            | 60,753               | 1,913,135            | —                 |
| Weighted average |                      |                      | 31.5              |

The reason for this difference between the simple and weighted arithmetic means will be readily understood. In the calculation of the simple mean in Table XLII each of the states is considered to be of equal importance. No account is taken of the fact that the individual average yields are associated with varying acreages and production in the several states. It is the weighted mean which properly adjusts these differences. Thus the low yield for North Dakota, in which state both the acreage and production are less than in any of the others, is not stressed beyond its own importance, while the high yield in Iowa, with an acreage and production that are greater than those of either of the other states, is given its full weight. The mean thus calculated in Table XLIII may be thought of as the simple arithmetic mean correctly calculated.

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 703, Table 46.

Another calculation of the weighted arithmetic mean is illustrated in Table XLIV.

TABLE XLIV. GRADES RECEIVED BY FIFTY-ONE SIXTH-GRADE PUPILS ON A SCIENCE TEST <sup>1</sup>

| 1                         | 2                                                         | 3                                                                           |
|---------------------------|-----------------------------------------------------------|-----------------------------------------------------------------------------|
| GRADE RECEIVED<br>(Class) | NUMBER OF PUPILS<br>RECEIVING THE<br>GRADE<br>(Frequency) | PRODUCT OF GRADE<br>AND NUMBER OF<br>PUPILS<br>(Column 1 times<br>Column 2) |
| 70                        | 3                                                         | 210                                                                         |
| 78                        | 6                                                         | 468                                                                         |
| 86                        | 10                                                        | 860                                                                         |
| 89                        | 13                                                        | 1,157                                                                       |
| 92                        | 9                                                         | 828                                                                         |
| 96                        | 7                                                         | 672                                                                         |
| 98                        | 2                                                         | 196                                                                         |
| 100                       | 1                                                         | 100                                                                         |
| Total      709            | 51                                                        | 4,491                                                                       |
| Weighted average   88.1   | —                                                         | —                                                                           |
| Simple average      88.6  | —                                                         | —                                                                           |

The weighted arithmetic mean is obtained by dividing the total of column 3 by the total of column 2. Hence, 4,491 divided by 51 gives 88.1, the weighted average grade received by the class of 51 pupils. The simple arithmetic mean of the 8 different grades in column 1 of Table XLIV is 88.6, obtained by merely summing the 8 grades and dividing by the number, 8. But this method of calculation does not take into account all the grades, but simply the 8 different grades. The weighted mean of 88.1, therefore, is in reality the simple arithmetic mean correctly calculated, since it is an average of all the grades and not merely an average of the 8 different grades received by the pupils.

<sup>1</sup> Selected from school records in Maryland.

From the foregoing calculations it will be observed that whenever different values are associated with unlike quantities the weighted arithmetic mean and the simple arithmetic mean will not be of the same magnitudes. On the other hand, however, when the different values are associated with like quantities the weighted and unweighted means will always be identical in size.

The calculation of a weighted average naturally involves the problem of sampling, and it is well to appreciate the fact that some of the uses that may be made of the weighted arithmetic average are of doubtful validity. It is no more than a truism to state that when all the factors of a group or universe are present in a sample in the same proportions as they exist in the group from which the sample was taken, the so-called weighted average and the simple average are of the same size. It often happens, of course, that when the universe is stratified it is not easily possible to select proportionate samples that show all the characteristic factors in the proper numbers or ratios. Hence, such a sample may easily be unrepresentative of the larger group from which it was taken, and in some such cases there may be justification for weighting the final results. The student will bear in mind as he progresses with his statistical work that he must not claim representativeness in every detail for a composite sample of data chosen and at the same time weight the results according to the association or ratio of varying values and quantities in the different parts of the sample. This may be inconsistent because, as has already been indicated, a strictly representative sample contains all the factors in the proper proportions, and for this reason the final average is weighted by ordinary calculation without the necessity of adjustments.

Perhaps this can be more easily understood by turning to Table XLII for an illustration. Here it is found that the unweighted average yield of corn per acre is 30.5 bushels, obtained by merely summing the yields in the several states and then dividing by the total number of individual yields. In Table XLIII are data pertaining to the same states represented in Table XLII, showing a weighted average yield of 31.5 bushels per acre. Now, the point to be made is that if a properly selected and truly representative sample is chosen from Table XLIII the weighted and unweighted average yields per acre as calculated from the sample will be identical. It is only because such truly representative samples are so difficult to select that there is ever any occasion for weighting the final results.

TABLE XLV. COSTS OF PRODUCING SPECIFIED QUANTITIES OF RUTABAGAS ON SELECTED FARMS <sup>1</sup>

| 1                               | 2                | 3                          | 4                        |
|---------------------------------|------------------|----------------------------|--------------------------|
| NUMBER OF FARMS                 | TOTAL PRODUCTION | COST OF PRODUCTION PER TON | TOTAL COST OF PRODUCTION |
|                                 | <i>Tons</i>      |                            |                          |
| 10                              | 100              | \$1.00                     | \$100.00                 |
| 20                              | 200              | 2.00                       | 400.00                   |
| 30                              | 300              | 3.00                       | 900.00                   |
| 40                              | 400              | 4.00                       | 1,600.00                 |
| 50                              | 500              | 5.00                       | 2,500.00                 |
| Total 150                       | 1,500            | 15.00                      | 5,500.00                 |
| Simple average cost per ton —   | —                | 3.67                       | —                        |
| Weighted average cost per ton — | —                | 3.67                       | —                        |

<sup>1</sup> Assumed data, selected for illustrative purposes only, and not at all for the purpose of conveying information as to actual costs. It will be observed that the number of farms producing at the higher costs increases as the costs increase. The assumption in this case is that there are other costs greater than \$5.00 per ton, with a decrease in the number of farms producing at costs greater than \$5.00.

Let us study the data in Table XLV. Here, we will assume, is a sample that is strictly representative of the group from which it was taken. Now, we shall see that the simple average of individual costs as shown in the table is the same as the weighted average cost. There are 150 farms, 10 of which show production costs to be \$1.00 per ton, 20 that show this cost to be \$2.00, and so on for the 30, 40, and 50 farms. There are, then, 150 separate costs, 10 of the \$1.00 costs, 20 of the \$2.00 costs, and so on. The total of these 150 individual costs per ton is \$550.00, which when divided by 150 gives \$3.67, the simple average cost.

The weighted average cost is obtained by dividing the total cost (summation of column 4) by the total production (summation of column 2). The result thus obtained

TABLE XLVI. FREQUENCY DISTRIBUTION OF FARM COSTS OF PRODUCING SORGHUM HAY PER TON IN GEORGIA AND TEXAS, 1925 <sup>1</sup>

| 1                                     | 2                                                                               | 3                                                                                         |
|---------------------------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|
| COST IN DOLLARS<br>PER TON<br>(Class) | FREQUENCY DISTRIBUTION<br>(Number of Tons Produced<br>at Varying Costs per Ton) | PRODUCT OF FREQUENCY<br>AND MIDPOINT OF CLASS<br>(Midpoint of Column 1 times<br>Column 2) |
| 1.51-2.00                             | 1                                                                               | 1.75                                                                                      |
| 2.01-2.50                             | 6                                                                               | 13.50                                                                                     |
| 2.51-3.00                             | 33                                                                              | 90.75                                                                                     |
| 3.01-3.50                             | 37                                                                              | 120.25                                                                                    |
| 3.51-4.00                             | 33                                                                              | 123.75                                                                                    |
| 4.01-4.50                             | 34                                                                              | 144.50                                                                                    |
| 4.51-5.00                             | 30                                                                              | 142.50                                                                                    |
| 5.01-5.50                             | 13                                                                              | 68.25                                                                                     |
| 5.51-6.00                             | 16                                                                              | 92.00                                                                                     |
| 6.01-6.50                             | 9                                                                               | 56.25                                                                                     |
| 6.51-7.00                             | 3                                                                               | 20.25                                                                                     |
| 7.01-7.50                             | 5                                                                               | 36.25                                                                                     |
| 7.51-8.00                             | 6                                                                               | 46.50                                                                                     |
| 8.01-8.50                             | 2                                                                               | 16.50                                                                                     |
| Total                                 | 228                                                                             | 973.00                                                                                    |
| Average cost per ton                  | —                                                                               | \$4.27                                                                                    |

<sup>1</sup> Costs as estimated.

is also \$3.67. We see, therefore, that when working with data of such nature as those shown in Table XLV there is nothing to be gained by weighting the final results.

### MEAN OF A FREQUENCY DISTRIBUTION

When the data are arrayed in the form of a frequency distribution the calculation of the mean, though a very simple process, is a little more detailed than the calculation of the mean of an ungrouped array or series. The method of calculation is illustrated in the preceding table.

In calculating the mean of a frequency distribution, particularly of a distribution in which the data are continuous, it is only necessary to multiply the midpoints of the classes by the number of frequencies in each class, summate the individual products, and then divide by the total number of frequencies in the distribution. Table XLVI illustrates how this is done. In column 1 are the classes. When the midpoints of the individual classes are multiplied by the frequencies in each respective class the products recorded in column 3 are obtained. The sum of column 3 is 973.00, which is divided by 228, the total of column 2, to obtain the true average of \$4.27 per ton.

By this method of calculation we assume the midpoints to be the true means of the classes, so that when the midpoint is multiplied by the number of frequencies the product is supposed to be the same as the product that would be obtained by multiplying each individual frequency by its own respective value and then summing the products. This assumption is often permissible in the case of continuous data, especially when there is a large number of items involved in each class. It has already been seen that by continuous data is meant such data as may have all conceivable fractional gradations. The



costs in Table XLVI may fall at any point in the class, so the frequencies are assumed to be arranged in such a way that the midpoints represent the actual average values. Therefore, by multiplying the midpoints of the classes by their respective frequencies and then summing the products a sum is obtained which when divided by the total number of frequencies gives the mean of the distribution. The reason for using the midpoints of the classes is that, assuming them to be the mean of the respective classes, the algebraic sum of the deviations is zero. The algebraic sum of the deviations measured from any point other than the mean would not give zero. If the items do not fall in such a distribution that the midpoints of the classes represent the true means, and if the weighted mean is to be calculated, it is necessary to multiply each individual of a class by its respective value, summate the individual products, and then divide the total by the total number of items. This will give a weighted mean of the distribution.

If the data be of the discrete type, as interest rates and farm mortgages, it is necessary to multiply each interest rate by the value of the mortgage, and then by the number of frequencies, summate, and divide by the total value of all the mortgages to obtain the mean. This would be the weighted mean of interest rates. The total sum obtained by multiplying interest rates and the number of frequencies would not be a product from which the true weighted mean could be calculated, since the magnitudes of the individual mortgages would not be given their proper weights. Seldom, if ever, can the mean of a discrete frequency distribution be calculated in the same way as the mean of a continuous frequency distribution. This is partly because it is not often statistically correct to

assume that a discrete series falls in a normal or other symmetrical distribution. There are exceptions to this, of course, and the problem is one that must be decided by the analyst himself, his judgment being based on the nature of the data with which he is dealing.

It is to be remembered by the student that since normal distributions seldom occur under actual conditions it is generally a better practice to calculate the mean by multiplying each individual item by its respective value, summing the products thus obtained, and then dividing by the total number of items. This, in fact, is almost always the only way by which the correct weighted mean of a frequency distribution can be obtained, but it is true that there are cases when the number of frequencies is so great that the most feasible way to calculate the mean is as illustrated in Table XLVI, page 105.

The mean of a frequency distribution can be calculated also from an assumed mean, by which method a correction is made for the difference between the true mean and the mean as it is assumed. This method of calculation is illustrated in Table XLVII.

In this case, for convenience, a mean is assumed and the deviations of the midpoints of the classes calculated. Then the deviations from the assumed mean are multiplied by the frequencies in each respective class, the products summated algebraically, and the total divided by the total number of frequencies. The quotient thus obtained is the mean deviation, an expression of the extent to which the true mean deviates from the assumed mean. This mean deviation may be either a plus or a minus quantity, depending upon the sign that accompanies the algebraic sum of the product of the frequencies and deviations from the assumed mean.

TABLE XLVII. FREQUENCY DISTRIBUTION OF FARM COSTS OF PRODUCING SORGHUM HAY PER TON IN GEORGIA AND TEXAS, 1925 <sup>1</sup>

| 1                                     | 2                                                                                        | 3                                                                               | 4                                                                                    |
|---------------------------------------|------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| COST IN DOLLARS<br>PER TON<br>(Class) | FREQUENCY<br>DISTRIBUTION<br>(Number of<br>Tons Produced<br>at Varying Costs<br>per Ton) | DEVIATION OF COST PER<br>TON FROM THE ASSUMED<br>MEAN<br>(Assumed Mean Is 4.25) | PRODUCT OF FRE-<br>QUENCY AND MID-<br>POINT OF CLASS<br>(Column 2 times<br>Column 3) |
| C                                     | f                                                                                        | d'                                                                              | fd'                                                                                  |
| 1.51-2.00                             | 1                                                                                        | -2.50                                                                           | -2.50                                                                                |
| 2.01-2.50                             | 6                                                                                        | -2.00                                                                           | -12.00                                                                               |
| 2.51-3.00                             | 33                                                                                       | -1.50                                                                           | -49.50                                                                               |
| 3.01-3.50                             | 37                                                                                       | -1.00                                                                           | -37.00                                                                               |
| 3.51-4.00                             | 33                                                                                       | -.50                                                                            | -16.50                                                                               |
| 4.01-4.50                             | 34                                                                                       | 0                                                                               | 0                                                                                    |
| 4.51-5.00                             | 30                                                                                       | +.50                                                                            | +15.00                                                                               |
| 5.01-5.50                             | 13                                                                                       | +1.00                                                                           | +13.00                                                                               |
| 5.51-6.00                             | 16                                                                                       | +1.50                                                                           | +24.00                                                                               |
| 6.01-6.50                             | 9                                                                                        | +2.00                                                                           | +18.00                                                                               |
| 6.51-7.00                             | 3                                                                                        | +2.50                                                                           | +7.50                                                                                |
| 7.01-7.50                             | 5                                                                                        | +3.00                                                                           | +15.00                                                                               |
| 7.51-8.00                             | 6                                                                                        | +3.50                                                                           | +21.00                                                                               |
| 8.01-8.50                             | 2                                                                                        | +4.00                                                                           | +8.00                                                                                |
| Total                                 | 228                                                                                      | —                                                                               | +4.00                                                                                |
| Mean cost<br>per ton                  | —                                                                                        | —                                                                               | \$4.27                                                                               |

In Table XLVII, \$4.25 is taken as the assumed mean, and the extent to which the midpoints of the classes in column 1 deviate from this assumed mean is given in column 3. As will be observed, the midpoints of classes that are smaller than the assumed mean carry a minus sign, and those that are larger carry a plus sign. The next step in the calculation of the true mean is to multiply the deviations in column 3 by the frequencies in column 2. The individual products thus obtained are recorded in column 4, the signs being carried. When all the products thus obtained have been entered in column 4 they are

<sup>1</sup> Costs as estimated.

summed algebraically and the resulting figure divided by the total number of frequencies in column 2 to obtain the average deviation. The summation of column 4 is 4.00, which divided by the total number of frequencies, 228, the total of column 2, gives .02, the average deviation, an expression of the extent to which the true mean deviates from the assumed mean. This quotient thus obtained is termed the correction factor. The assumed mean is \$4.25, and the true mean is obtained by adding .02, which gives \$4.27. If the correction factor had been a minus quantity it would have been subtracted from the assumed mean to obtain the true mean.

### PROGRESSIVE MEAN

The progressive mean is merely a series of simple arithmetic averages. It is obtained by including an additional item each time an average is to be calculated. Suppose, for example, we have data showing the quantity of rice imported into this country during each month of a specified year, beginning with January and ending with December. The progressive mean would be calculated by dividing the total quantity for January and February by 2, the total for January, February, and March by 3, and so on, adding one month each time.

### GEOMETRIC AND HARMONIC MEANS

There are at least two other means that are sometimes used in statistical analysis which may be mentioned. The geometric mean, for example, is the  $n$ th root of the product of the factors. Thus, when there are only two items or numbers the geometric mean is the square root of their product. The geometric mean of a series of three items is the cube root of their product, and so on for other series

of greater numbers of items. Its calculation is facilitated by the use of logarithms, which consists merely of determining the arithmetic average of the logarithms of each item or number in the series. The quotient thus obtained is the logarithm of the geometric mean. It may be interesting to observe that the geometric mean of a series is always less than the simple arithmetic mean, and that it is obviously zero when any item or number in the series is zero.

The harmonic mean is the reciprocal of the arithmetic average of the reciprocals of the items or numbers in a series. Its calculation, then, consists merely of summing the reciprocals of the items, dividing the summation by the number of reciprocals, and determining the reciprocal of the quotient.<sup>1</sup>

<sup>1</sup> For a discussion of the uses and application of the geometric and harmonic means see E. E. Day, *Statistical Analysis*, 1927, Chapter X.

## CHAPTER VI

### THE MODE

The mode is simply the item or value which occurs the greatest number of times, or, in other words, it is the point of maximum frequency, coinciding with the arithmetic mean in a normal distribution. When the size of each and every item in an array or series is known the mode may be located by simply making a count of how many times each one occurs. In Table XLVIII, for example, a count of the several grades shows that 92 occurs more times than any other grade. It is the grade of 92, therefore, that is called the mode or modal grade.

TABLE XLVIII. GRADES MADE BY TEN NINTH-GRADE PUPILS ON A BIOLOGY TEST <sup>1</sup>

| <i>Grade</i> |
|--------------|
| 90           |
| 91           |
| 92           |
| 92           |
| 92           |
| 93           |
| 94           |
| 95           |
| 95           |
| 96           |

The mode is more readily located if the items are arranged in a frequency distribution. The following table shows this very clearly. It is immediately seen that 92 occurs more times than any other value.

<sup>1</sup> Specially selected from Maryland school records.



TABLE XLIX. FREQUENCY DISTRIBUTION OF GRADES MADE BY TEN NINTH-GRADE PUPILS ON A BIOLOGY TEST <sup>1</sup>

| GRADE<br>(Class) | FREQUENCY<br>(Number of Times the<br>Grade Occurred) |
|------------------|------------------------------------------------------|
| C                | f                                                    |
| 90               | 1                                                    |
| 91               | 1                                                    |
| 92               | 3                                                    |
| 93               | 1                                                    |
| 94               | 1                                                    |
| 95               | 2                                                    |
| 96               | 1                                                    |
| Total            | 10                                                   |

TABLE L. FREQUENCY DISTRIBUTION OF AGES OF INDIVIDUALS <sup>2</sup>

| AGE IN YEARS<br>(Class) | FREQUENCY<br>(Number of Individuals<br>Whose Ages Fall within<br>Specified Ranges) |
|-------------------------|------------------------------------------------------------------------------------|
| C                       | f                                                                                  |
| 1-5                     | 30                                                                                 |
| 6-10                    | 41                                                                                 |
| 11-15                   | 60                                                                                 |
| 16-20                   | 68                                                                                 |
| 21-25                   | 73                                                                                 |
| 26-30                   | 91                                                                                 |
| 31-35                   | 105                                                                                |
| 36-40                   | 116                                                                                |
| 41-45                   | 128                                                                                |
| 46-50                   | 135                                                                                |
| 51-55                   | 87                                                                                 |
| 56-60                   | 85                                                                                 |
| 61-65                   | 61                                                                                 |
| 66-70                   | 48                                                                                 |
| 71-75                   | 39                                                                                 |
| 76-80                   | 20                                                                                 |
| Total                   | 1,187                                                                              |

<sup>1</sup> Based on data in Table XLVIII, page 112.<sup>2</sup> Specially selected data.

When continuous data are arranged by groups in a frequency distribution it is only the modal class that can be readily located by mere inspection. The preceding table will serve to illustrate this fact, and also to show how the mode may be calculated.

We see at once that 46-50 is the modal class, since the number of individuals with ages falling within those limits is greater than the number falling within the limits of any other class. The so-called true mode, therefore, is located somewhere within the interval between 46 and 50. A hasty approximation of the mode could be made by assuming it to be located at the midpoint of the class. For purposes of more nearly accurate calculation, however, the following formula has been devised.<sup>1</sup>

$$Mo = L + \frac{f_2 c}{f_2 + f_1}$$

In this formula the meanings of the symbolic expressions are as follows:

$Mo$  = the mode

$L$  = the lower limit of the class in which the mode is located

$f_2$  = the number of frequencies in the next higher class, or in the class immediately above the one in which the mode is located

$c$  = the class interval

$f_1$  = the number of frequencies in the next lower class, or in the class immediately below the one in which the mode is located.

In Table L it is seen that in the class immediately preceding the one in which the mode is located there are more frequencies than in the class immediately following.

<sup>1</sup> W. I. King, *Elements of Statistical Method*, 1923, page 124.

It is to be expected, therefore, that the true mode will fall nearer to the lower limit of the modal class than to the upper limit. The calculation of the mode will show this to be true. With the formula we can now proceed to locate the theoretically true mode. By substituting the proper values for the symbolic expressions in the formula we have

$$Mo = 46 + \frac{87 \times 5}{87 \times 128} = 46.04,$$

the required mode.

The degree of accuracy of the results obtained by the application of this formula is greatest when used in connection with continuous data which fall or tend to fall in a normal frequency distribution.

Calculation of the mode of a sample of data naturally involves the question of whether or not it is applicable to the universe from which the sample data were selected. This, of course, depends upon the degree of representativeness of the sample selected. If the sample truly represents the universe from which the individual items were selected, then the mode of the sample will coincide with the mode of the total group. On the other hand, however, if there is some doubt as to the representativeness of the sample the data may be adjusted by smoothing the frequency curve to obtain a better representation of the universe, in which case the calculated mode would in reality be an estimated mode, probably very different from the true mode of the entire group.

An important feature of calculation of the mode by this method, and which may invalidate the results obtained, is that sometimes there may be no actual value that corresponds to the mode as located or determined. It is for this reason that whenever the number of items is

not so large as to make it impracticable the mode is more satisfactorily located by arranging the items in an array according to their size, rather than by grouping them into classes with designated limits. This is particularly true of a discrete series.

It is often the modal value that is referred to when in common parlance we speak of the "average." We say, for example, that in certain areas of North Carolina the average farmer plants 20 acres of cotton, meaning, of course, that there are more acreages of this size than there are of any other.

## CHAPTER VII

### THE MEDIAN

The median is the item which is located in the middle of an array, dividing the array into two equal parts so far as number of items is concerned. That is to say, the median is that item which has the same number of items preceding it as there are following, and, like the mode, it coincides with the arithmetic mean in a normal distribution. In the following table are shown the statistics of acre yields of tobacco in the United States for five years.

TABLE LI. YIELD OF TOBACCO PER ACRE IN THE UNITED STATES, 1924-1928 <sup>1</sup>

| YEAR | YIELD PER ACRE |
|------|----------------|
|      | <i>Pounds</i>  |
| 1924 | 733.6          |
| 1925 | 783.3          |
| 1926 | 783.6          |
| 1927 | 764.7          |
| 1928 | 718.3          |

TABLE LII. ARRAY OF TOBACCO YIELDS IN THE UNITED STATES FOR FIVE YEARS, 1924-1928 <sup>2</sup>

| RANK | YIELD PER ACRE |
|------|----------------|
|      | <i>Pounds</i>  |
| 1    | 718.3          |
| 2    | 733.6          |
| 3    | 764.7          |
| 4    | 783.3          |
| 5    | 783.6          |

<sup>1</sup> *Yearbook, United States Department of Agriculture, 1928*, page 893, Table 341.

<sup>2</sup> Based on data in Table LI above.

To locate the median yield we array the items in Table LI in their respective order of ascending magnitude. We could just as well, of course, array the yields in descending order of magnitude.

Recalling that the median is so located that there are as many items preceding as there are following, its determination becomes an exceedingly simple matter. We see that there are 5 items in the array, and that the third item has 2 items preceding it and 2 following. The third item, 764.7, therefore, is the median. If there were 4 items in the array, the first four items, say, under "Yield per acre," the median would fall midway between the second and third items. Since there is no value corresponding to that location, we would assume the median to be the average of the second and third items. The calculated median, then, would be 749.15. If the second and third items were of the same magnitude the median, it will be seen, would coincide with them.

Perhaps the location of the median of an array will be facilitated by increasing the number of items by 1 and then dividing by 2. For example, in Table LII there are 5 items. Increasing this number by 1 gives 6. When 6 is divided by 2 a quotient of 3 is obtained, meaning that the third item in the array is the median. Assuming there were only 4 items in the array, the quotient obtained by adding 1 to 4 and then dividing by 2 would be  $2\frac{1}{2}$ . This would mean that the median as calculated is located midway between the second and third items in the array.

When the items are arranged in a frequency distribution it is only the class in which the median is located that can be determined by a mere count. The following table is presented for illustrative purposes.



TABLE LIII. FREQUENCY DISTRIBUTION OF THE YIELD OF RYE PER ACRE IN THE UNITED STATES FOR NINETEEN YEARS, 1910-1928 <sup>1</sup>

| YIELD PER ACRE IN BUSHELS<br>(Class) | FREQUENCY<br>(Number of Times the<br>Yield Occurred) |
|--------------------------------------|------------------------------------------------------|
| C                                    | f                                                    |
| 11.0-11.9                            | 2                                                    |
| 12.0-12.9                            | 3                                                    |
| 13.0-13.9                            | 2                                                    |
| 14.0-14.9                            | 2                                                    |
| 15.0-15.9                            | 5                                                    |
| 16.0-16.9                            | 4                                                    |
| 17.0-17.9                            | 1                                                    |
| Total                                | 19                                                   |

There are 19 items or frequencies in the table above. According to the method that has been outlined the median is the tenth item, located by adding 1 to 19 and then dividing by 2. The class in which the median is located may be determined by counting the frequencies downward from the top or upward from the bottom until the tenth item is reached. Counting from the top, we summate 2, 3, 2, and 2, which gives a total of 9, but the tenth item is the median. We know, therefore, that the median is located somewhere in the 15.0-15.9 class. A mere count will not show us, however, what the numerical value of the median is. We do not know, for instance, whether the median in Table LIII is 15.1, 15.5, or some other value. For definitely locating the median in such frequency distributions the following formula has been found quite convenient: <sup>2</sup>

$$M = L + \frac{c(2i-1)}{2f}.$$

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 693, Table 33.

<sup>2</sup> W. I. King, *Elements of Statistical Method*, 1923, page 129.

The meanings of the symbolic expressions contained in this formula are as follows:

$M$  = the median

$L$  = the lower limit of the class in which the median is located

$c$  = the interval of the class containing the median

$i$  = the number of items that must be counted in the median class to determine where the median occurs

$f$  = the number of items in the class in which the median occurs.

Substituting numerical values for the symbolic expressions in the formula we obtain

$$M = 15 + \frac{1(2 \times 1 - 1)}{2 \times 5} = 15.1, \text{ the required median.}$$

Perhaps additional explanation of the substitutions will make the calculation more fully understood. The figure 15 is "L," the lower limit of the 15.0-15.9 class in Table LIII. The figure 1 before the brackets is "c," the class interval of the 15.0-15.9 class. Within the brackets we have, first,  $2 \times 1$ , which is 2 times the number of items that must be counted in the 15.0-15.9 class to locate the median. We have already seen that the frequencies of the 4 classes preceding give a total of 9, indicating that 1 additional item must be counted in the next class. Below the line is  $2 \times 5$ , which is simply 2 times the number of frequencies in the 15.0-15.9 class.

A most important feature of this formula is the fact that its application is based on the assumption that the items are distributed in the class uniformly according to size. In Table LIII, therefore, since it has been determined that the median falls at 15.1, the assumption is

that the other 4 items in the 15.0–15.9 class fall at 15.3, 15.5, 15.7, and 15.9, or in some similar distribution. Whenever the items of a class do not vary uniformly in size the median as calculated by means of the formula is only the approximate median.

## CHAPTER VIII

### VARIABILITY AND DISPERSION

In the discussion of the arithmetic mean, the mode, and the median an attempt was made to show how a measure or expression of central tendency may sometimes be used to convey an impression of characteristics of an entire group of items. The arithmetic mean, for example, is an expression of the average size of items, or the point from which the algebraic sum of the deviations is zero. The mode tells us what value or what item occurs the greatest number of times, while the median is simply that item which is centrally located in an array, with the same number of items preceding it as there are following. These expressions of central tendency, however, do not indicate the extent to which items vary from them in magnitude or degree. The arithmetic mean, we see, does not tell us the size of the smallest, largest, or any other item in the group from which the mean was calculated. We know only that some of the items are smaller than the mean and that some are larger. The mode merely shows which value or item occurs more times than any other. It may be the first item, or the last, or any other in the group. The median tells us only the size of the item that is located at the midpoint of an array. It does not indicate definitely the size of items preceding or following. In statistical analysis it becomes necessary, therefore, in expressing variation to adopt some measure other than that of central tendency to adequately show the degree to which items deviate

from a central location. A few of the more commonly used of these measures will be explained and illustrated.

### FREQUENCY TABLE

It has already been shown how items may be arranged in a frequency table according to their respective magnitudes. When a group of items is thus arranged by classes in order of size of individuals there is furnished a rather clear conception of the nature of the data considered as a whole. We observe in Table XLVI,<sup>1</sup> for example, that the number of tons of hay produced are arrayed in a frequency distribution according to cost of production. Selecting any frequency we choose, it can be readily seen what the associated cost is. Furthermore, the frequency table shows the magnitude of items both below and above any designated location. Suppose in Table XLVI we single out the maximum frequency of 37 in column 2. It can be seen at once that of the 228 tons of hay the production costs of 37 tons fall within a range of from \$3.01 to \$3.50. In addition to this, by mere addition it is readily determined that of the 228 tons the production costs of 40 tons were less than for the 37 tons, and that the costs of producing 151 tons were greater. The frequency table provides, therefore, a rather definite basis for measuring variability in magnitudes and values.

### FREQUENCY GRAPH

When a frequency distribution is graphically presented the number of items associated with a specific value can be read directly, and at the same time the relative position of each class readily determined. Let us turn to the histogram in Chart 23.<sup>2</sup> Here it is immediately seen that 3 pupils live within 1 block of their school, 7 live between 1

<sup>1</sup> Page 105.

<sup>2</sup> Page 87.

and 2 blocks away, and so on. Thus, any group of pupils can be singled out from the others and its position in relation to other groups immediately seen. Take, for instance, the group of nine pupils who live between 3 and 4 blocks away from the school. We see at once not only their location as to distance from school, but also that there are three groups of pupils comprising 17 individuals who live nearer the school, and that there are three other groups including a total of 14 pupils who live farther from school than either of the individuals in the group of nine.

### THE RANGE

When we know the smallest and largest items in an array or series, we have some indication of the characteristics of the group to which these items belong. In Table LII the smallest item is seen to be 718.3, whereas the largest item has a value of 783.6, showing a range between the two of 65.3. It is known, therefore, that the other values in the series can be no smaller than 718.3 and no larger than 783.6. Obviously, then, the remaining items fall within a range of 718.3 plus 65.3, or within a range of 783.6 minus 65.3. Likewise, in Table LIII, a frequency distribution arranged in classes, it is readily seen that the 19 items fall within the lower limits of the first class and the upper limits of the last class, or, in other words, all the items fall within a range of 6.9. These illustrations should serve to show that the range is an indication of dispersion, or of the general nature of variability in the distribution.

### MEAN DEVIATION OF A SERIES

The extent to which items deviate on an average from a designated central tendency furnishes a quite satisfactory measure of dispersion in many instances. This average



deviation, ordinarily calculated from some such tendency, may be based on measures other than the simple arithmetic mean, so that in expressing the numerical value of the mean deviation it is essential that the particular measure of central tendency be stated.

As the name implies, the mean deviation is merely the simple average of deviations, signs being ignored. Its calculation consists simply of determining the mean difference between the measure of central tendency and the individual items comprising the group. Table LIV illustrates calculation of the mean deviation from the simple arithmetic mean.

TABLE LIV. MONTHLY MEAN TEMPERATURE AT MONTGOMERY, ALABAMA <sup>1</sup>

| MONTH                                 | MONTHLY MEAN<br>TEMPERATURE IN<br>FAHRENHEIT DEGREES | DEVIATIONS FROM<br>THE ANNUAL AVERAGE<br>(Signs Ignored) |
|---------------------------------------|------------------------------------------------------|----------------------------------------------------------|
| January                               | 48.2                                                 | 17.3                                                     |
| February                              | 51.6                                                 | 13.9                                                     |
| March                                 | 57.8                                                 | 7.7                                                      |
| April                                 | 65.3                                                 | .2                                                       |
| May                                   | 73.4                                                 | 7.9                                                      |
| June                                  | 79.6                                                 | 14.1                                                     |
| July                                  | 81.7                                                 | 16.2                                                     |
| August                                | 80.8                                                 | 15.3                                                     |
| September                             | 76.3                                                 | 10.8                                                     |
| October                               | 66.6                                                 | 1.1                                                      |
| November                              | 55.8                                                 | 9.7                                                      |
| December                              | 49.4                                                 | 16.1                                                     |
| Annual average                        | 65.5                                                 | —                                                        |
| Mean deviation from<br>annual average | —                                                    | 10.8                                                     |

The mean deviation of monthly temperature from the annual average is found to be 10.8 degrees, indicating that on an average the monthly mean temperature deviates 10.8 degrees from the mean for the entire year. The

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 135, Table 146.

mean deviation could just as well be calculated from the median or mode of these monthly temperatures, or even from the figure as given for any one month. We could, for instance, base the mean deviation on the temperature for January, determining to what extent, on an average, the temperatures of the other months deviate. The important feature to remember, as has been stated, is that in expressing the mean deviation the measure from which it is calculated must always be given.

### MEAN DEVIATION OF A FREQUENCY DISTRIBUTION

When the data are arranged in a frequency distribution, or in the form of a frequency table, the mean deviation

TABLE LV. FREQUENCY DISTRIBUTION OF THE YIELD OF RYE PER ACRE IN THE UNITED STATES FOR NINETEEN YEARS, 1910-1928 <sup>1</sup>

| 1                                       | 2                                                    | 3                                                                                                    | 4                                                                                                                           |
|-----------------------------------------|------------------------------------------------------|------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| YIELD PER ACRE<br>IN BUSHELS<br>(Class) | FREQUENCY<br>(Number of Times<br>the Yield Occurred) | DEVIATION OF THE<br>MIDPOINT OF EACH<br>CLASS FROM THE<br>MEAN<br>(Signs Ignored)<br>Mean = 14.605 * | PRODUCT OF DEVIATION OF THE MIDPOINT OF EACH CLASS FROM THE MEAN AND THE NUMBER OF FREQUENCIES<br>(Column 3 times Column 2) |
| C                                       | f                                                    | d'                                                                                                   | fd'                                                                                                                         |
| 11.0-11.9                               | 2                                                    | 3.105                                                                                                | 6.210                                                                                                                       |
| 12.0-12.9                               | 3                                                    | 2.105                                                                                                | 6.315                                                                                                                       |
| 13.0-13.9                               | 2                                                    | 1.105                                                                                                | 2.210                                                                                                                       |
| 14.0-14.9                               | 2                                                    | .105                                                                                                 | .210                                                                                                                        |
| 15.0-15.9                               | 5                                                    | .895                                                                                                 | 4.275                                                                                                                       |
| 16.0-16.9                               | 4                                                    | 1.895                                                                                                | 7.580                                                                                                                       |
| 17.0-17.9                               | 1                                                    | 2.895                                                                                                | 2.895                                                                                                                       |
| Total                                   | 19                                                   | —                                                                                                    | 29.695                                                                                                                      |
| Mean deviation                          | —                                                    | —                                                                                                    | 1.573                                                                                                                       |

\* The mean is obtained by multiplying the frequencies of each class by the midpoints of the classes, summing, and dividing by the number of frequencies. If the deviations in column 3 are multiplied by the frequencies and summated algebraically the product will be zero.

<sup>1</sup> Data from Table LIII, page 119.

tion is calculated by determining the difference between the mean and the midpoints of each class, signs ignored, multiplying this difference in each instance by the number of frequencies in the particular class, and then summing and dividing by the total number of frequencies in the distribution. Table LV illustrates the calculation of the mean deviation from a frequency distribution.

The mean of the frequency distribution in Table LV is 14.605. Column 3 shows the deviations of the midpoints of the classes from the mean, and in column 4 are recorded the products obtained by multiplying the deviations in column 3 by the frequencies in column 2. Summing these products, we obtain a total of 29.695, which when divided by 19, the number of frequencies, gives 1.573, the mean deviation of the frequency distribution. This mean deviation is in fact in terms of ranges, since the data are arranged in classes with definite intervals. This interval is 1, meaning there is a range of 1 between the midpoints of the classes. The mean deviation, then, in terms of bushels is the product obtained by multiplying the mean deviation in terms of ranges by 1. In Table LV these two mean deviation values coincide, and this calculation is not necessary. If, however, the range had been greater or less than 1 the mean deviation in terms of ranges would have been multiplied by the range to obtain the mean deviation in terms of actual values.

The mean deviation of 1.573 as calculated signifies that this is the average difference in size of items in Table LV. It is obvious, of course, that in those cases where the mean deviation is small the indications point conclusively to the fact that the items tend to center about the arithmetic average more than if the mean deviation is large. An important feature of the mean deviation as calculated from

a frequency distribution is that it is based on the assumption of even distribution in the several classes, with the midpoint of each class or group being selected as the representative item. If the distribution is normal the mean deviation based on the arithmetic average is about .79798 as large as the standard deviation, discussed later in this chapter. As implied, however, and as is so obvious, this relationship between the mean deviation and the standard deviation is not maintained for skewed distributions.

### COEFFICIENT OF DISPERSION

This measure is useful in comparing the dispersion of one group of data with that of another. We may have two normal distributions of exactly the same number of frequencies, with the same average deviation as calculated from the arithmetic mean, mode, or median, and yet there may be a difference in the relative dispersion of the two groups. The following hypothetical data are presented for illustrative purposes.

| GROUP 1 |           | GROUP 2 |           |
|---------|-----------|---------|-----------|
| COST    | FREQUENCY | COST    | FREQUENCY |
| \$0     | 1         | \$1     | 1         |
| 1       | 10        | 2       | 10        |
| 2       | 45        | 3       | 45        |
| 3       | 120       | 4       | 120       |
| 4       | 210       | 5       | 210       |
| 5 Mean  | 252       | 6 Mean  | 252       |
| 6       | 210       | 7       | 210       |
| 7       | 120       | 8       | 120       |
| 8       | 45        | 9       | 45        |
| 9       | 10        | 10      | 10        |
| 10      | 1         | 11      | 1         |
| Total   | 1,024     | —       | 1,024     |

There is the same number of items or frequencies in each group, and, of course, the deviations from the respective means are identical. The average of deviations from the mean of group 1, therefore, is the same as the corresponding average deviation of group 2. But the means are different in magnitude, which, even though average deviations are equal, indicates a difference in relative dispersion. The coefficient of dispersion of these average deviations is obtained by simply dividing the average deviation by the arithmetic mean. Thus we have

$$\frac{\text{Average deviation}}{\text{Average}} = \text{Coefficient of dispersion.}^1$$

It will be seen that when average deviations of two groups of data are identical the group with the smaller measure of central tendency from which the average deviation was calculated will have the larger coefficient of dispersion. In the preceding data the average deviation of both group 1 and group 2 as measured from the means is 1.23, obtained by dividing the summation of the deviations from the mean by 1,024, the number of deviations. Since the mean of group 1 is less than the mean of group 2, it will be immediately seen that the result is larger when the average deviation, 1.23, is divided by 5 than when divided by 6. This so-called result is the coefficient of dispersion, which in this case is merely the average deviation expressed as a fractional part of the mean.<sup>1</sup> When the coefficients of dispersion for two groups of data are unlike it is an indication that relative deviations are of different magnitudes. Comparisons of this kind, of course, must be based on average deviations calculated from the same measures of central tendency.

<sup>1</sup> Multiplying this fraction by 100 converts the coefficient of dispersion to a percentage.

## STANDARD DEVIATION

The standard deviation, a measure of dispersion from the arithmetic mean,<sup>1</sup> is calculated by squaring the deviation of each item from the mean, summing the squares, dividing by the number of observations, and then extracting the square root. This measure of dispersion has been called the "root-mean-square deviation," which, as indicated, is merely the square root of the average of the squares of the deviations from the simple arithmetic average. The calculation of the standard deviation of a simple array is illustrated in Table LVI.

TABLE LVI. GRADES MADE BY FIVE NINTH-GRADE PUPILS ON A BIOLOGY TEST <sup>2</sup>

(Calculation of the Standard Deviation of an Array) \*

| 1                | 2                                                       | 3                                                         |
|------------------|---------------------------------------------------------|-----------------------------------------------------------|
| GRADE<br>(Class) | DEVIATION FROM<br>SIMPLE ARITHMETIC<br>MEAN (Mean = 92) | SQUARE OF<br>DEVIATIONS FROM<br>SIMPLE ARITHMETIC<br>MEAN |
| C                | d                                                       | d <sup>2</sup>                                            |
| 90               | -2                                                      | 4                                                         |
| 91               | -1                                                      | 1                                                         |
| 92               | 0                                                       | 0                                                         |
| 93               | 1                                                       | 1                                                         |
| 94               | 2                                                       | 4                                                         |
| Total            | —                                                       | 10                                                        |

\* The standard deviation of a series, such as cotton production for the period from 1900 to 1930, is calculated in the same way.

Proceeding with the calculation we have the following:

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{10}{5}} = 1.41, \text{ in which}$$

<sup>1</sup> Sometimes referred to as a quadratic mean, which is the square root of the arithmetic average of the squares of the items or numbers in a series. It is in this sense, therefore, that the standard deviation is the quadratic mean of the deviations from the arithmetic mean. The standard deviation, of course, may be calculated from any measure of central tendency. In common usage, however, it is seldom based on any average but the simple arithmetic mean.

<sup>2</sup> Data selected from Table XLVIII, page 112.



$\Sigma d^2$  is the summation of the squares of the deviations from the simple arithmetic mean, as represented by the total of the products in column 3, and  $N$ , the number of items or observations.<sup>1</sup>

The standard deviation, as has been stated, is a measure of the extent to which items in a series deviate from the simple arithmetic mean of the series, giving weight, as indicated, to extreme deviations. If in a normal-frequency distribution we measure off on both sides of the mean a distance equal to the standard deviation we have an area or range that includes approximately 68.26 per cent. of the total number of items in the distribution.<sup>2</sup> Let us assume the mean of a normal distribution of 200 items to be 20 and the standard deviation to be 6. Measuring off a distance of 6 to both the right and left of the arithmetic mean, we have an area including about 68.26 per cent. of the total 200 items. The distance thus measured from both sides of the mean extends to the points on a normal-frequency curve where the curve changes from a concave to a convex surface, or to the point of inflection. The standard deviation of an accurate and truly representative sample is the same as the standard deviation of the entire group of data, and even if the individual samples are not representative there is a tendency for the curve of aver-

<sup>1</sup> The calculation of the standard deviation may be more readily comprehended if it is realized that the term "items or observations" in this case may be construed to mean that it refers in reality to the number of deviations from the arithmetic mean. It is obvious, of course, that the number of original items in a series and the number of deviations from the mean are identical. This applies also, of course, to frequency distributions. Then, too, the standard deviation is really calculated from the deviations from the mean, so that the student may prefer to let  $N$  equal the number of deviations.

<sup>2</sup> This fact has been worked out by mathematicians and can be definitely proved by exact measurement. Table XC, Appendix A, shows the areas of the probability surface that are included within the range of .01 to 5.00 standard deviations when measured from the arithmetic mean of a normal distribution.

ages to be normal, though, of course, the standard deviations of individual samples not representative are different from the standard deviation of the entire group.

The standard deviation calculated from a frequency distribution is illustrated in the following table.

TABLE LVII. FREQUENCY DISTRIBUTION OF THE HEIGHTS OF INDIVIDUALS <sup>1</sup>  
(Calculation of the Standard Deviation from a Frequency Distribution)

| 1                                              | 2                                                                 | 3                                                                                                | 4                                                                                            | 5                                                                                                              |
|------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| HEIGHT OF<br>INDIVIDUALS<br>IN FEET<br>(Class) | FREQUENCY<br>(Number of<br>Individuals<br>of Specified<br>Height) | DEVIATION<br>FROM THE MEAN*<br>(Difference<br>between the<br>Mean and<br>Column 1)<br>(Mean = 5) | PRODUCT OF<br>FREQUENCY<br>AND DEVIATION<br>FROM THE MEAN<br>(Column 2<br>Times<br>Column 3) | FREQUENCY<br>TIMES THE<br>SQUARE OF<br>DEVIATION FROM<br>THE MEAN<br>(Column 2 Times<br>Square of<br>Column 3) |
| C                                              | f                                                                 | d <sup>1</sup>                                                                                   | fd <sup>1</sup>                                                                              | f(d <sup>1</sup> ) <sup>2</sup>                                                                                |
| 3                                              | 1                                                                 | -2                                                                                               | -2                                                                                           | 4                                                                                                              |
| 4                                              | 2                                                                 | -1                                                                                               | -2                                                                                           | 2                                                                                                              |
| 5                                              | 3                                                                 | 0                                                                                                | 0                                                                                            | 0                                                                                                              |
| 6                                              | 4                                                                 | 1                                                                                                | 4                                                                                            | 4                                                                                                              |
| Total                                          | 10                                                                | *                                                                                                | 0                                                                                            | 10                                                                                                             |

\* If the deviations from the mean are multiplied by the frequencies and the products summated algebraically the total will be zero.

The mean in the preceding table is calculated by multiplying the heights of individuals by the corresponding number of frequencies (column 1 by column 2), summing, and dividing by 10, the total number of frequencies. The summation of column 5 represents the total of the products obtained by multiplying the square of deviations from the mean by the number of frequencies. It is this summation, therefore, that is in reality the total of the squares of deviations from the mean of the frequencies. If we array the frequencies in column 2 in order of magnitude we would have the following:

<sup>1</sup> Hypothetical data.

3  
4  
4  
5  
5  
5  
6  
6  
6  
6

There is one frequency with a size of 3, two with a size of 4, three with a size of 5, and four with a size of 6. The summation of the squares of deviations of each of the above values from the mean of 5 is 10, the same as in column 5 of Table LVII. Calculation of the summation of deviations squared is facilitated by the method in the table when a large number of deviations is involved. Obviously, if there are 100 frequencies, say, distributed among 5 classes it would ordinarily be much easier to calculate the total of the deviations from the mean squared in accordance with Table LVII than by arraying the 100 items, determining the deviation of each item from the mean, squaring these deviations, and then summing.

In connection with this explanation it may be mentioned that when the deviations from the mean of a frequency distribution are multiplied by the corresponding frequencies and the products summated algebraically the total or result will always be zero. This is of the same significance as saying that the algebraic sum of the deviations from the arithmetic mean is zero. Column 4 of Table LVII illustrates this fact.

The summation of column 5, Table LVII, which, as has been indicated, is simply the summation of the products

of the frequencies times the squares of the deviations from the mean, is the value from which the standard deviation is calculated.<sup>1</sup> Thus we have the following:

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\sum d^2}{N}} = \sqrt{\frac{10}{10}} = 1.0.$$

In this calculation the class interval is 1, as shown by the difference in height in column 1 of Table LVII, meaning that the standard deviation in terms of both class intervals and feet is the same. If the class interval were greater or less than 1, it would be necessary to multiply the standard deviation in terms of class intervals by the

TABLE LVIII. FREQUENCY DISTRIBUTION OF HEIGHTS OF INDIVIDUALS <sup>2</sup>  
(Calculation of the Standard Deviation in Terms of Class Interval Units from an Assumed Mean and Converting the Standard Deviation to Actual Values)

| 1                                     | 2                                                     | 3                                                                                                                                                 | 4                                                                                  | 5                                                                                                 |
|---------------------------------------|-------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| HEIGHT OF INDIVIDUALS IN FEET (Class) | FREQUENCY (Number of Individuals of Specified Height) | DEVIATION FROM THE ASSUMED MEAN, OR NUMBER OF CLASS INTERVAL UNITS THE MIDPOINT OF EACH CLASS DEVIATES FROM THE ASSUMED MEAN (Assumed mean = 5.7) | PRODUCT OF FREQUENCY AND DEVIATION FROM THE ASSUMED MEAN (Column 2 Times Column 3) | FREQUENCY TIMES THE SQUARE OF DEVIATION FROM THE ASSUMED MEAN (Column 2 Times Square of Column 3) |
| C                                     | f                                                     | d <sup>1</sup>                                                                                                                                    | fd <sup>1</sup>                                                                    | f(d <sup>1</sup> ) <sup>2</sup>                                                                   |
| 5.0-5.2                               | 1                                                     | -3                                                                                                                                                | -3                                                                                 | 9                                                                                                 |
| 5.2-5.4                               | 2                                                     | -2                                                                                                                                                | -4                                                                                 | 8                                                                                                 |
| 5.4-5.6                               | 12                                                    | -1                                                                                                                                                | -12                                                                                | 12                                                                                                |
| 5.6-5.8                               | 17                                                    | 0                                                                                                                                                 | 0                                                                                  | 0                                                                                                 |
| 5.8-6.0                               | 9                                                     | 1                                                                                                                                                 | 9                                                                                  | 9                                                                                                 |
| 6.0-6.2                               | 7                                                     | 2                                                                                                                                                 | 14                                                                                 | 28                                                                                                |
| 6.2-6.4                               | 2                                                     | 3                                                                                                                                                 | 6                                                                                  | 18                                                                                                |
| Total                                 | 50                                                    | 0                                                                                                                                                 | 10                                                                                 | 84                                                                                                |

<sup>1</sup> It is obvious that this summation is merely the total of the squares of all the deviations from the arithmetic mean.

<sup>2</sup> Hypothetical data.

magnitude of the class interval in order to have it expressed in terms of feet.

The preceding table illustrates the method of calculating the standard deviation in terms of class interval units from an assumed mean and then converting this standard deviation to terms of actual values.

In the preceding table the assumed mean is 5.7, and it deviates  $\frac{10}{50}$ , or 0.2 of a class interval unit from the true

mean. A correction must be made, therefore, in order to obtain the true mean. Since 0.2 is a plus quantity, the true mean is 0.2 of a class interval unit greater than the assumed mean. A class interval unit in Table LVIII is 0.2 of a foot, and when this 0.2 is multiplied by the number of class intervals, or the fractional part of a class interval, that the true mean deviates from the assumed mean (0.2 times 0.2) the product of 0.04, the correction factor, is obtained. Adding 0.04 to the assumed mean, 5.7, we obtain 5.74, the true mean of height of individuals. If the assumed mean had been the correct mean the summation of  $fd^1$  (column 4) would have been zero, as in the case of column 4 of Table LVII.

The standard deviation may now be calculated as follows:

$$\sigma = \sqrt{\frac{\Sigma f(d^1)^2}{N} - C^2} = \sqrt{\frac{84}{50} - .04} = 1.28.^1$$

This value of 1.28 is the standard deviation in class interval units, in which  $\Sigma f(d^1)^2$  is the sum of the products obtained by multiplying the squares of the deviations from

<sup>1</sup> The formula may be written also as  $\sigma = \sqrt{\frac{\Sigma f(d^1)^2 - C^2 \times N}{N}}$ , in which  $C^2 \times N$

is the correction factor squared times the number of items, the number of items being the same as the number of deviations from the assumed mean.



the assumed mean by the number of frequencies,  $N$ , the number of items or observations (which may be interpreted also as the number of deviations from the assumed mean of the frequencies), and  $C$ , the correction factor in class interval units. To convert this standard deviation to terms of feet we simply multiply it by the size of the class interval. We have already seen the class interval to be 0.2 of a foot. Therefore, multiplying 1.28 by 0.2 we obtain .256, the standard deviation in feet. If the standard deviation had been calculated in feet it could have been converted to class interval units by dividing it by the magnitude of the class interval.

Perhaps it will be well to note the fundamental principles underlying the calculations based on Table LVIII. The assumed mean, as we have seen, was 5.7. We know this is not the true mean because the algebraic summation of column 4, which is the sum of the deviations from the assumed mean, is greater than zero. The total summation of column 4 is 10, which is expressed in class interval units, since the deviations from the assumed mean, column 3, are in terms of class intervals. From this it follows that when the deviations in class interval units are multiplied by corresponding frequencies in column 2 and the products summated algebraically, the summation, likewise, is in terms of class intervals. Therefore, dividing the total of deviations, 10, by 50, the number of items, a value of 0.2, a plus quantity, is obtained, which also, of course, is in class interval units. This value of 0.2 is the correction factor, which must be added to the assumed mean to obtain the true mean. But class interval units cannot be added to feet, so it becomes necessary to convert the correction factor to feet, the term in which the assumed mean is expressed. In column 1 of Table LVIII



it is seen that a class interval unit is 0.2 feet. Since 0.2 of a class interval unit must be added to the assumed mean we simply take 0.2 of 0.2 of a foot and add it to 5.7, which gives 5.74, the true mean.

The same correction factor would be obtained by arraying the 50 items according to size of the midpoints of the classes, determining the deviations in class interval units of these midpoints from the assumed mean, squaring these deviations, summing, and dividing by the number of deviations, which would be 50, the same as the number of frequencies or items. We see, then, that by such an arrangement there are 19 minus deviations and 29 plus deviations. The difference between these is 10, corresponding to the summation of column 4 in Table LVIII, and the number of deviations, which is always the same as the number of items or frequencies, is 50. This method, like that followed in the table, gives the correction factor in terms of class interval units because all the deviations are in terms of these units. If the deviations had been expressed in feet, the correction factor as calculated would have been on that basis also.

The standard deviation has been seen to be the value obtained by extracting the square root of the average of the squares of the deviations from the mean. In Table LVIII the total of column 5 is 84, the summation of the squares of the deviations of frequencies from the assumed mean. The average of this summation of squares is obtained by dividing by 50, the number. Before extracting the square root, however, it is necessary to correct for the true mean, so that we will have the average of the squares of deviations from the true mean. The principle can be illustrated by repeating the formula for the standard deviation:

$$\sigma = \sqrt{\frac{\Sigma d^2 - C^2}{N}}.$$

The value of  $C^2$  is the square of the correction factor in terms of class interval units. It will be readily seen why this standard deviation is in these terms. The deviations in column 3 are in units of the class interval, and when the items in column 2 are multiplied by corresponding items in column 3 the products thus obtained are also in terms of class interval units. It follows, then, that the total of column 5 is the summation of the deviations in class interval units squared. The correction factor as calculated from the algebraic sum of deviations from the assumed mean in class interval units (total of column 4) being also a value in terms of class intervals, the standard deviation must necessarily be in the same terms.

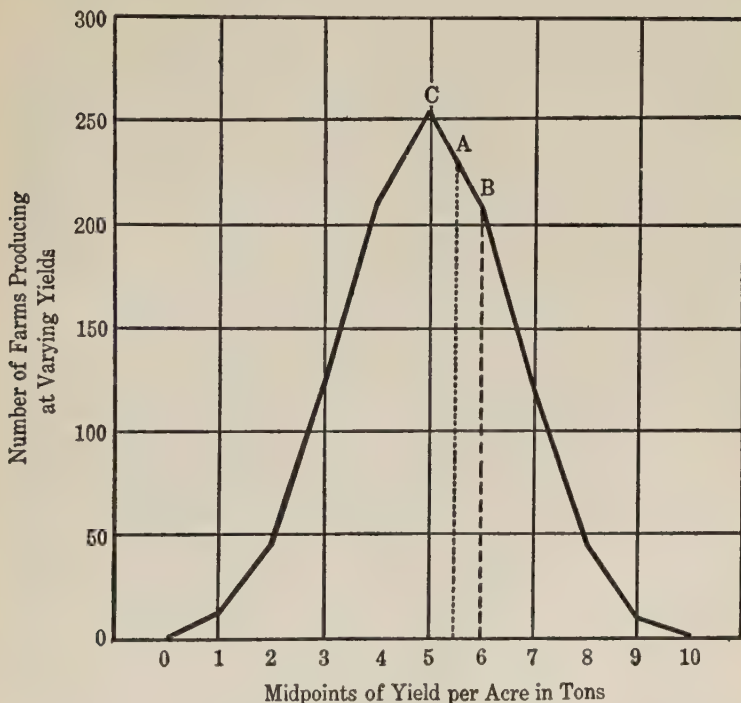
TABLE LIX.    NORMAL-FREQUENCY DISTRIBUTION OF THE YIELD  
OF TOMATOES PER ACRE <sup>1</sup>

| YIELD PER ACRE<br>IN TONS<br>(Class) | FREQUENCY<br>(Number of Farms<br>Producing at<br>Varying Yields) |
|--------------------------------------|------------------------------------------------------------------|
| C                                    | f                                                                |
| 0                                    | 1                                                                |
| 1                                    | 10                                                               |
| 2                                    | 45                                                               |
| 3                                    | 120                                                              |
| 4                                    | 210                                                              |
| 5                                    | 252                                                              |
| 6                                    | 210                                                              |
| 7                                    | 120                                                              |
| 8                                    | 45                                                               |
| 9                                    | 10                                                               |
| 10                                   | 1                                                                |
| Total                                | 1,024                                                            |

<sup>1</sup> Hypothetical data, based on the binomial expansion to the 10th power.

CHART 28. NORMAL-FREQUENCY DISTRIBUTION OF THE YIELD OF  
TOMATOES PER ACRE

(Hypothetical data, based on the binomial expansion to the 10th power, as shown  
in Table LIX)



The calculation of a standard deviation from a frequency distribution such as that in Table LVIII is based on the assumption that the frequencies are uniformly distributed in the several classes, which may or may not be true.

When the standard deviation is calculated from the midpoints of the classes in a normal-frequency distribution in which the class intervals are uniform, a correction

should be made for the true measure. This is because the midpoints of classes as measured in linear distance on the  $x$  axis do not divide the number of frequencies equally. The principle is illustrated in Chart 28, which is based on Table LIX.

The chart, purposely left unsmoothed, represents a total of 1,024 items, 252 of which are associated with the  $x$  value of 5. Since 5 is the midpoint of a class, these 252 items are distributed equally about the midpoint designated as "C." There are 210 items associated with the  $x$  value of 6. In the same way that the  $y$  values are associated with the  $x$  value of 5, these 210 items are distributed equally around the point designated as "B." The dotted line "A" is drawn vertically on the  $y$  axis from the midpoint of the linear distance between the  $x$  values of 5 and 6. This line, however, does not divide the total number of items lying between "C" and "B" into two equal parts, because, as has been indicated, there are more items between "C" and the midpoint of the linear distance between the values of 5 and 6 than there are between the midpoint of this linear distance and line "B." It follows, therefore, that the line equally dividing the total number of items between "C" and "B" lies a little to the left of line "A," or, in other words, a little closer to "C" than to "B." By this illustration we may quite readily see that in such frequency distributions a linear measurement on the  $x$  axis that equally divides the  $x$  values does not necessarily divide the  $y$  values into two equal parts. The dividing point to the left of line "A" is so located that the parallelograms on either side of it are of equal area. When the linear distance between the  $x$  values of 5 and 6 is equally divided and a line drawn vertically from the midpoint of this distance to the point of intersection with the

curve, we have two parallelograms of different areas. This is illustrated in Chart 28. Line "A" extends vertically from the midpoint of the values of 5 and 6 as measured in linear distance on the x axis. It is readily seen that the area between lines "C" and "A" is greater than the area between "B" and "A."

The preceding explanation will suffice to show that when a standard deviation is calculated from the midpoints of classes in a normal-frequency distribution, a correction is necessary because the midpoint of the class is represented by linear distance on the x axis. The midpoints of linear distances between classes, then, excepting the modal class of the entire distribution, do not coincide with the point that equally divides the number of items in the classes of normal distributions, and yet this is the assumption when the standard deviation is calculated from a normal distribution without a correction.

A very convenient formula has been devised for making this correction.<sup>1</sup> It is as follows:

Standard deviation ( $\sigma$ ) =  $\sigma^2$  (as calculated) -  $\frac{C^2}{12}$ , in which

$C$  is the class interval, and  $\frac{1}{12}$  a fractional expression

applying to normal distributions. For any other distribution some other fraction would be used. The calculations involved in this formula are very simple. The first step is to square the standard deviation as calculated from the deviations of the midpoints of the classes from the mean. Then the class interval is squared, the product divided by 12, and the result subtracted from the square

<sup>1</sup> Commonly known as Sheppard's formula. See Mills, F. C., *Statistical Methods*, pages 530 and 537, and Yule, G. U., *Introduction to the Theory of Statistics*, pages 212 and 307.

of the standard deviation as originally calculated from the squares of the deviations of the midpoints of classes from the mean. This gives the square of the corrected standard deviation. Extracting the square root, we obtain the required standard deviation in actual values.

An important feature of the standard deviation is that it varies in size according to both the number of items and the deviations from the arithmetic mean. A large summation of squares of deviations from the mean may be offset by a large number of deviations. On the other hand, there may be such large deviations that the standard deviation is greater than the arithmetic mean, in which case it is of no value as a measure of dispersion. The reason the standard deviation of a truly representative sample is the same as the standard deviation of the group from which the sample was taken is that the means are identical, with deviations from the means that correspond both as to magnitudes and proportionate numbers. Obviously, then, if we calculate the arithmetic mean of a truly representative sample and of the entire universe from which the sample was taken the two means will be found to be of the same magnitudes, and if the deviations of items in each group from their respective means are determined, these deviations of different sizes will be in evidence in each series of deviations in exactly the same proportions. Unless this latter is true the sample is not truly representative, even though its standard deviation be of the same size as the standard deviation of the whole universe. When all the items in a series are of the same size it is obvious that the standard deviation is zero.

The formula for the standard deviation of any frequency distribution following the normal law of error may be expressed as  $\sigma = \sqrt{NPQ}$ . The meanings of these



expressions may be made clear by making reference to the tossing of coins. In this case,  $N$  would be the number of coins tossed,  $P$  would be the probability of obtaining a head in the toss of each coin, and  $Q$  would be the probability of not obtaining a head in the toss of a coin. Let us assume 10 coins are being tossed. The formula for the standard deviation would then be  $\sigma = \sqrt{10 \times \frac{1}{2} \times \frac{1}{2}} = 1.58$ . If 20 coins are being tossed the square of the standard deviation will be twice as big as the square of 1.58, but the standard deviation itself will not be twice as big as 1.58.

It is likely that in the analysis of many economic data the average deviation will be considered preferable to the standard deviation, unless for some reason special weight is to be given to extreme variations from the mean. A deviation of 2, for instance, from the arithmetic mean is 1/6 as great as a deviation of 12, whereas the square of a deviation of 2 is only 1/36 as large as the square of a deviation of 12.

| 1            | 2                         | 3            | 4                         | 5                                       | 6                                                                    | 7                                                                               |
|--------------|---------------------------|--------------|---------------------------|-----------------------------------------|----------------------------------------------------------------------|---------------------------------------------------------------------------------|
| a            | da<br>(dev. from<br>mean) | b            | db<br>(dev. from<br>mean) | a - b<br>(column<br>1 -<br>column<br>3) | (a - b) -<br>mean of<br>a - b<br>(column 5<br>- mean of<br>column 5) | Square of<br>deviations<br>of a - b from<br>its mean<br>(square of<br>column 6) |
| 18           | 2                         | 14           | 0                         | 4                                       | 1.8                                                                  | 3.24                                                                            |
| 16           | 0                         | 13           | -1                        | 3                                       | .8                                                                   | .64                                                                             |
| 22           | 6                         | 17           | 3                         | 5                                       | 2.8                                                                  | 7.84                                                                            |
| 20           | 4                         | 19           | 5                         | 1                                       | -1.2                                                                 | 1.44                                                                            |
| 15           | -1                        | 15           | 1                         | 0                                       | -2.2                                                                 | 4.84                                                                            |
| 19           | 3                         | 14           | 0                         | 5                                       | 2.8                                                                  | 7.84                                                                            |
| 15           | -1                        | 15           | 1                         | 0                                       | -2.2                                                                 | 4.84                                                                            |
| 14           | -2                        | 13           | -1                        | 1                                       | -1.2                                                                 | 1.44                                                                            |
| 12           | -4                        | 10           | -4                        | 2                                       | -.2                                                                  | .04                                                                             |
| 9            | -7                        | 10           | -4                        | 1                                       | -1.2                                                                 | 1.44                                                                            |
| Mean<br>= 16 | 0                         | Mean<br>= 14 | 0                         | Mean<br>= 2.2                           | 0                                                                    | Total<br>= 33.60                                                                |

In analyzing two samples of similar data it may be necessary in calculating the standard deviation to make a correction for the differences between them and for the smallness of sample. To illustrate these principles let us refer to the preceding hypothetical data.

Column 1 represents the first series and column 3 represents the second series. The deviations of the individual items of each series from their means are recorded in columns 2 and 4 respectively. Column 5 shows the differences between pairs of corresponding items in columns 1 and 3. The deviations of individual items in column 5 from their mean, the calculation of which mean removes the correlation between the series, are recorded in column 6, and column 7 contains the square of these deviations. The standard deviation of the differences between the two series is obtained from the summation of column 7 by the common method of calculation. This standard deviation is found to be 1.83 (square root of one-tenth of 33.60).

To correct for the standard deviation of the entire universe from which the samples were selected, we ascertain how big the standard deviation of the sample should be in relation to the standard deviation of the entire universe. Consulting *Tables for Statisticians and Biometricians*<sup>1</sup> we find that for a sample of 10 the standard deviation should be .8944 as big as the standard deviation of the entire group from which the sample of 10 was chosen. Proceeding with the correction, it is but necessary to divide the standard deviation as calculated, 1.83, by .8944. This gives a quotient of 2.05, the standard deviation as corrected for the entire universe.

<sup>1</sup> W. F. Sheppard, Cambridge University Press, edited by Karl Pearson.

## STANDARD DEVIATION FROM THE MEDIAN

In calculating the standard deviation from a median the procedure is identical with that used in the calculation based on deviations from the mean, except that the standard deviation from a median cannot be based on deviations from an assumed median. All deviations, whether from the median of an array or from the median of a frequency distribution, must be determined from the true median.

## COEFFICIENT OF VARIATION

When the standard deviation is divided by the simple arithmetic average and the product multiplied by 100 we have the standard deviation expressed as a percentage of the average. This measure is known as the coefficient of variation. The formula for its calculation is as follows:

$$\text{Coefficient of variation} = \frac{\text{Standard deviation}}{\text{Simple arithmetic average}} \times 100.$$

As will be remembered, once a group of data has been chosen the size of the standard deviation depends upon the magnitude of the squares of the deviations from the arithmetic mean and the number of items in the group. It is a natural conclusion, therefore, that the size of the coefficient of variation is an indication of the extent to which items deviate from their mean. If, for example, the coefficient of variation is 50 per cent. as large as the mean the implications are that a distance equal to this amount measured off on both sides of the mean will include approximately 68.26 per cent. of the total number of items in the group from which the mean was calculated. This principle always applies to normal distributions, and there is a tendency for it to apply to other distributions because

the curve of averages follows, or tends to follow, the normal curve of distribution. We see, therefore, that the smaller the coefficient of variation, the greater is the tendency for items to group themselves about the simple arithmetic mean.

### QUARTILE

The discussion relating to the median will aid in explaining the meaning and the method of locating the quartiles. It will be recalled that the median is that item which divides an array into two equal parts, so located that there are as many items preceding it as there are following. The quartiles, then, may be defined as the items, values, or magnitudes that divide an array into four equal parts. Obviously, there are three quartiles, the second one coinciding with the median.<sup>1</sup> The first quartile of a simple array of items may be located by merely increasing the number of items in the array by 1 and then dividing by 4. The second quartile (the median) may be located by increasing the number of items by 1, multiplying by 2, and then dividing by 4. The third quartile may be located by increasing the number of items by 1, multiplying by 3, and then dividing by 4. We have, therefore, the following formulas for locating the three quartiles:

$$\text{First quartile} = \frac{n+1}{4}$$

$$\text{Second quartile} = \frac{2(n+1)}{4}$$

$$\text{Third quartile} = \frac{3(n+1)}{4}, \quad \text{in which } n$$

equals the number of items in the array.

<sup>1</sup> The median may correctly be called the second quartile, but in common usage it is not generally referred to as such.

The second quartile may be located also by multiplying the location of the first quartile by 2, and the third quartile may be located by multiplying the location of the first quartile by 3.

To illustrate the application of the formulas that have been given for the location of the quartiles, let us select an array of 11 items as follows:

CALCULATION OF THE QUARTILES OF AN ARRAY IN WHICH  
THERE IS AN ODD NUMBER OF ITEMS

| <u>Array</u>  |                          |
|---------------|--------------------------|
| 10            |                          |
| 11            |                          |
| 12            | First quartile           |
| 13            |                          |
| 14            |                          |
| 15            | Second quartile (median) |
| 16            |                          |
| 17            |                          |
| 18            | Third quartile           |
| 19            |                          |
| 20            |                          |
| <u>      </u> |                          |
| $n=11$        |                          |

The position of the first quartile of this array is located by means of the formula  $\frac{n+1}{4}$ . Substituting, we have

$\frac{11+1}{4}=3$ , the position of the first quartile, coinciding with the value of 12 in the array. The position of the second quartile is located by means of the formula  $\frac{2(n+1)}{4}$ ,

which by substitution  $=\frac{2(11+1)}{4}=6$ . The sixth item in the array, therefore, which is 15, is the second quartile. The position of the third quartile is located by means of the formula  $\frac{3(n+1)}{4}$ , which, with the proper substitu-

tions,  $= \frac{3(11+1)}{4} = 9$ . The ninth item in the array, then,

which is 18, represents the position of the third quartile.

When there is an even number of items in a simple array the positions of the quartiles are located in the same manner as when the array contains an odd number of items. Accordingly, in a simple array of 12 items the first quartile would fall at a position located one-fourth of the distance between the third and fourth items. The second quartile would fall at a position located at one-half of the distance between the sixth and seventh items, and the third quartile would fall at a position located at three-fourths of the distance between the ninth and tenth items. In actual practice, of course, the positions of the

TABLE LX. FREQUENCY DISTRIBUTION OF THE AVERAGE ACREAGE OF TOMATOES IN NINETEEN MARYLAND COUNTIES GROWN FOR CANNING PURPOSES <sup>1</sup>

(Calculation of Quartiles)

| 1                          | 2                                                                                     |
|----------------------------|---------------------------------------------------------------------------------------|
| NUMBER OF ACRES<br>(Class) | FREQUENCY<br>(Number of Counties with<br>Acreages Falling within<br>Specified Ranges) |
| C                          | f                                                                                     |
| 0-999                      | 8 First quartile<br>= 562                                                             |
| 1,000-1,999                | 1                                                                                     |
| 2,000-2,999                | 4 Median = 2,125                                                                      |
| 3,000-3,999                | 3 Third quartile<br>= 3,500                                                           |
| 4,000-4,999                | 1                                                                                     |
| 5,000-5,999                | 1                                                                                     |
| 6,000-6,999                | 1                                                                                     |
| Total                      | 19                                                                                    |

<sup>1</sup> Based on records of the Maryland Agricultural Experiment Station at College Park.



quartiles thus located may not coincide with any given value in the array. It is the graphic location, however, that is of most importance in discussing method of finding the quartile positions.

Calculation of the approximate location of the quartiles in a frequency distribution, as in a simple array, is not at all a complicated procedure. Several methods have been used in locating the position of the so-called precise quartiles of a frequency distribution, all of which have their merits. A simple method, satisfactory for nearly all practical purposes, will be illustrated by using the data in the preceding frequency table.

The median of this distribution located by means of the formula  $L + \frac{c(2i-1)}{2f}$ , as presented in Chapter VII, is 2,125, indicating that the number of counties with acres less than 2,125 is the same as the number with acres that are greater. This value corresponds to the tenth frequency in column 2. There are, then, 9 items in the distribution on either side of the median. Considering each of these groups of 9 items as entirely separate, we proceed to calculate the medians. In the first group the median is 562, and in the second group it is 3,500. These medians are the calculated quartiles of the total distribution, 562 representing the first, and 3,500 the third. We see by this calculation that the first quartile is the fifth item in the frequency column, the median, the tenth item, and that the third quartile coincides with the fifteenth item. It does not necessarily follow, of course, that the quartiles calculated from a frequency distribution coincide with the quartiles as calculated from an array of the same data. The reasons for this are too obvious for further discussion.

## QUARTILE DEVIATION

This measure of dispersion is simply the quantity corresponding to one-half the difference between the first and third quartiles. It is sometimes referred to in statistical analysis as the semi-interquartile range, one-half of the range between lower and upper quartiles. Calculation consists of subtracting the first quartile from the third and dividing by 2. If, for example, the first (lower) quartile is 10 and the third (upper) quartile is 16, we have a semi-interquartile range, or quartile deviation, of 3, obtained by subtracting 10 from 16 and then dividing by 2.

It will be seen, then, that the quartile deviation is the average deviation of the quartiles from the median in an array. In the preceding illustration we know that the median is so located that the respective quartiles are the third items above and below it. From this it will be seen also that if the items in an array are condensed around the median value the distance between quartiles will be relatively small, with a corresponding small quartile deviation. If the items do not tend to center around the median the first and third quartiles will be farther away, meaning that the quartile deviation will be greater.

The important feature to remember about the quartile deviation is that 50 per cent. of the total number of items in an array is included in the area lying between the first and third quartiles. Therefore, when the magnitudes of the quartiles and the distance between them are known we have a rather definite indication of the nature of the data, and, since the median lies midway between the first and third quartiles, we know that 25 per cent. of the total number of items lies between it and either quartile.

## DECILES AND PERCENTILES

As implied, deciles divide an array or distribution into ten equal parts, whereas percentiles make a hundred divisions. These measures of dispersion are seldom used in ordinary statistical analysis. Calculation is by methods that have already been illustrated for the quartiles and median, and for this reason no additional explanation is necessary.

## SKEWNESS <sup>1</sup>

Skewness is merely distorted normal distribution. It will be recalled that in a normal-frequency curve the mean, mode, and median coincide. That is to say, if the mean of a normal distribution is 10, it is a certainty that 10 is the item that occurs more times than any other, and that has just as many items preceding it as there are following. A normal distribution, therefore, has a skewness of zero.<sup>2</sup>

There are two kinds of skewness, negative and positive. If in a curve or distribution the mode falls nearer the right extreme than to the left extreme the distribution is said to be negatively skewed. On the other hand, if the greater part of the curve lies to the right of the mode the distribution possesses positive skewness. It will be seen, therefore, that the mean may be either greater or less than the mode, depending upon the nature of skewness, and that the median value may lie somewhere between the mode and mean. Since, then, skewness is an indication that the mean, mode, and median do not coincide it is possible to

<sup>1</sup> See *Asymmetrical Distribution*, Chapter III, pages 63 and 64.

<sup>2</sup> A normal distribution has a kurtosis of 3, which may be calculated by dividing the first moment to the fourth power by the square of the second moment.

$$(\text{Kurtosis}) B_2 = \frac{M^4}{M_2^2}$$

obtain a measure that shows this deviation from normal distribution, or from any other distribution. One of the most familiar methods consists of subtracting the mode from the mean and then dividing by the standard deviation. This quotient or value thus obtained is an expression of how many standard deviations, or what part of a standard deviation, the mean deviates from the mode. Whenever the mode is greater than the mean there is a minus quantity when it is subtracted from the mean, indicating that the distribution is negatively skewed. When the mean is greater than the mode there is positive skewness. By merely determining the difference between the mean and mode without subsequent division by any measure we have an expression in actual values that serves as an indication of skewness. If, for example, the value of 6 is the mean and the value of 4 is the mode it can be seen immediately that the mode is 2 less than the mean.

There are numerous other methods of measuring skewness that have been formulated, but since they are of extremely doubtful value they are not being referred to in this text. Certainly, the method outlined here is practicable, and it is this feature of statistical technique that we wish to stress.

### PROBABLE ERROR

Another measure which may be mentioned at this time is the probable error, which may be termed an expression of reliability.<sup>1</sup> It is a numerical quantity such that the chance error of an observation in a universe is equally likely to be of greater or less magnitude. As applied to the arithmetic mean, for example, it is the median of the deviations. If we were to calculate the deviations of

<sup>1</sup> Sometimes interpreted as a measure of unreliability.

individual items in a normal distribution from their mean, and then arrange these deviations in an array according to their magnitude without regard to whether they are plus or minus quantities, the median of these deviations, the item which divides the array of deviations into two equal parts, would be the so-called probable error. For instance, if the arithmetic mean is 25 and the probable error, the median deviation, is 6, then it is obvious that one-half of the items deviate more than 6 from the arithmetic mean, whereas the other one-half of the total number of items deviate less than 6 from the mean. The chances are equal, therefore, that any item selected at random from the entire group is as likely to fall within the range of the average plus or minus the probable error, 6, as it is to fall outside of this range.

In statistical analysis the probable error is used as a measure of reliability in many instances, and the interpretation is always fundamentally the same for all constants. The essential feature in regard to its precise calculation is based on the assumption that the distribution is normal. Whenever this is not the case the probable error does not convey its characteristic significance in full degree. In the following chapter a few of the more common applications of the probable error will be explained and illustrated.



## CHAPTER IX

### PROBABILITY AND ERROR

The preceding discussion regarding measures of dispersion has already indicated the nature of probability. Let us inquire a little further into the exact nature of its meaning.

#### NORMAL-FREQUENCY CURVE

The normal-frequency curve represents probable occurrences in an infinite number of observations or trials. As suggested, it is known also as the probability curve, and often is referred to as the curve of error and as the Gaussian curve, after Karl Friedrich Gauss, German mathematician and physicist, who, among many accomplishments, first worked out a solution for binomial equations, and to whom much of the earlier analysis of the theory of probability is accredited. In discussing the standard deviation, we saw that in a normal distribution when a distance equal to the standard deviation is measured off on either side of the mean we have an area that includes approximately 68.26 per cent. of the total number of items. The distance from the mean to the quartile is equal to .6745 (or 67.45 per cent.) of the standard deviation, and when measured from both sides of the mean we include 50 per cent. of all the items in the distribution.

Under actual conditions normal-frequency distributions of many types of economic data seldom occur. In farm costs of producing a certain commodity, for example,



a random sample of individual costs is likely to form a skewed distribution, and often, if not in most cases, this skewness will be positive. The reason for this latter tendency becomes apparent when it is realized that although farm costs can never be as low as zero they may conceivably reach any limit on the upper scale, with the result that in a large number of instances the modal value falls nearer the left end, or closer to the origin, of the curve of distribution than it does to the other extremity. This, of course, is not intended as an implication that the mode of such a sample will always occur near the point of origin of the curve of distribution, since the inherent character of the data may be such that the proportion of lesser values is small. Many conditions and circumstances may enter into a particular situation to offset what is often a long-time tendency, so that in a particular instance a sample distribution selected at random from a universe in which there is a time element may exhibit negative skewness in spite of the fact that over a long period of time the group from which the sample was taken is characteristically positively skewed. It is a fairly safe assertion to make, however, that a representative sample of economic data will possess positive skewness in the vast majority of cases. A normal-frequency distribution of such data, and particularly of farm costs, is quite indicative of unrepresentativeness, and its reliability as a basis for measuring the characteristics of the group from which selected becomes a matter of speculation. Furthermore, whenever the total number of individual costs of a sample of farm cost data fall in a normal distribution, it is very likely that the sampling process has not included the proper proportions of high and low costs. It is a natural conclusion, therefore, that in interpreting the standard deviation of a

distribution it is to be considered as being an exponent of normal-frequency.

#### APPLICATION OF THE PROBABLE ERROR

The probable error, as has been indicated elsewhere, is that quantity or amount which in any number of observations is less than one-half the errors and greater than the other half. We have already seen that in a normal-frequency distribution if a distance equal to .6745 of the standard deviation is measured off from both sides of the arithmetic mean we have an area or range including 50 per cent. of all the items in the distribution. This numerical value, then, is the probable error of the distribution, since any item selected at random is equally likely to be larger or smaller than the mean plus or minus .6745 of the standard deviation. It is true also that one-half of such selected items are as likely to be within the range of the mean minus .6745 of the standard deviation as they are to be within the range of the mean plus .6745 of the standard deviation. The probable error, then, coincides with the middle or median item of deviations arrayed without regard to whether they are plus or minus quantities.

The probable error may be calculated for and applied to all the various constant measures used in statistical analysis, and for the coefficients as well. A few of the more common of these applications will be briefly explained.

#### *Probable Error of the Mean of a Sample*

The probable error of the arithmetic mean of a sample is calculated by dividing the square root of the number of items into the standard deviation and then multiplying by .6745. The formula for this calculation is as follows:

$$\text{Probable error of the mean (P.E.)} = .6745 \frac{\sigma}{\sqrt{N}} \cdot$$

The second part of this formula shows us how big the square root of the average of the squares of the deviations from the mean of the sample is in comparison to the size of the square root of the number of items themselves. But the probable error of the mean on which the standard deviation is based is only .6745 as great as this quotient or quantity because it is in effect a quartile deviation.

If we have a distribution of 100 items with a mean of 50 and a standard deviation of 20, the probable error of the mean is 1.349. This indicates that when arithmetic means are calculated from other random selections of 100 items selected on the same basis as the first sample, the chances are equal that one-half of these means will fall within a range of 50, the mean of the first group, plus or minus 1.349. In other words, the means of subsequently selected samples are as likely to fall between the range of 48.651 (50 minus 1.349) and 51.349 (50 plus 1.349) as they are to be less than 48.651 or greater than 51.349. In an entire universe, particularly if the distribution is normal, the probable error of the mean may be interpreted as the limits on either side of the arithmetic mean within which a random selection of an individual is as likely to occur as it is to occur outside these limits.

The probable error of the mean of a sample may be quite small and yet a bias in sampling may make it unrepresentative. For instance, if we have an error of 1 it simply indicates that 50 per cent. of the means of an infinite number of samples of data similar to the first sample are likely to fall within a range of the mean of the first sample plus or minus 1. It does not necessarily indicate

that the original sample as selected is representative of the entire group or universe.

A problem may arise from time to time as to what really constitutes the number of items of which the square root is to be extracted in calculating the probable error of the arithmetic mean. Let us assume we have a sample of cost records on peanuts representing 100 farms and 1,000,000 pounds. The question is, what is  $N$ , the 100 farms or the 1,000,000 pounds? It will be seen how easily this is decided. There are 100 farms, with costs calculated for all of them. There are, then, 100 of these individual average costs, and it is this number that is used in the calculation of the standard deviation, and which, accordingly, must be used in determining the probable error. The reader will appreciate the fact that costs were obtained for each of the 100 individual farms, and not for each of the 1,000,000 pounds of peanuts. Obviously, it would not only be impracticable, but impossible as well, to determine the cost of producing each of the 1,000,000 pounds. The probable error of the mean of a sample like this indicates the variation in costs on individual farms, and not the variation in costs of producing single units of one pound.

The probable error of the mean is quite commonly used as a measure of reliability. Whenever the probable error is no greater than one-sixth of the standard deviation the sample of data may be considered fairly reliable, and when it is no greater than one-twelfth of the standard deviation it is a safe deduction that the sample is sufficiently reliable for practical purposes. In chance or random selections the probabilities involved in sampling are such that in only one case in a hundred is there a chance of the arithmetic average being inaccurate to the extent of more than four times the size of the associated probable error. That is, if

we were to select a sample at random and then calculate the mean and the probable error of the mean there is but one chance in a hundred that the true mean of the entire universe would differ from the mean of the sample to the extent of more than four times the probable error of the mean of the sample.

### *Probable Error of a Single Variate*

The probable error of a single variate is the quartile distance of the curve of individual variates. Any single observation subsequently made at random is as likely to fall within the range of the quartiles as it is to be less than the first quartile or greater than the third quartile. The probable error of a single variate may be calculated, therefore, by the same formula as the probable error of the mean. Accordingly, a probable error of the mean of 1.349 indicates that any individual variate chosen at random is as likely to fall within a range of the mean plus or minus 1.349 as it is to fall outside of this range. It indicates also that when a number of single variates is chosen at random the chances are even that one-half of those falling within the range of the mean plus or minus 1.349 will be greater than the mean and one-half will be less.

### *Probable Error of a Standard Deviation*

When we divide the standard deviation by the square root of two times the number of items and then multiply by .6745 the probable error of the standard deviation is obtained. Let us assume a distribution or series of 50 items with a standard deviation of 5. The square root of 100 (two times 50) is 10, which when divided into 5, the standard deviation, gives a quotient of .5, and when .5 is multiplied by .6745 a product of .33725 is obtained. It is



this value that is the probable error of the standard deviation. For convenience the following formula has been devised to facilitate the calculation.

Probable error of the standard deviation

$$(P.E.\sigma) = 0.6745 \frac{\sigma}{\sqrt{2N}}.$$

The probable error of the standard deviation of a sample indicates that there is an even chance that the standard deviation of the universe lies somewhere within the range of the standard deviation of the sample plus or minus its probable error. Hence, with a standard deviation of 5 and a probable error of .33725 it is a safe assumption that the true standard deviation is as likely to lie within a range of 5 plus or minus .33725 as it is to be outside this range. A further indication is that if a random selection is made from an infinite number of standard deviations calculated from samples similar to the first the chances are even that the standard deviation selected will fall within the limits of the first, plus or minus its probable error.

#### *Probable Error of the Coefficient of Correlation*

The formula for the probable error of the coefficient of correlation is  $.6745 \frac{(1-r^2)}{\sqrt{N}}$ . Calculation by this formula

consists of squaring the coefficient of correlation,<sup>1</sup> subtracting the product from 1, dividing by the square root of the number of pairs of items, and then multiplying by .6745. Let us assume a coefficient ( $r$ ) of .90 and a series of 100 pairs of items. The square of .90 is .81, which sub-

<sup>1</sup> Correlation is discussed in a later chapter. In the formula,  $r$  represents the coefficient of correlation. The probable errors of the correlation ratio and the correlation index are calculated from the same general formula, with the proper measures being substituted for  $r$ .



tracted from 1 leaves .19. Dividing .19 by 10, the square root of 100, gives .019. Multiplying .019 by .6745 gives .013, the probable error of the coefficient of correlation.

The significance of this error is that one-half of any other coefficients of correlation calculated from samples of data similar to the first, will, on the average, fall within a range of the first coefficient plus or minus .013, assuming random selection in each instance. Of those falling within this range there is an even chance that any one selected at random is as likely to be larger than the first as it is to be smaller.

### *Probable Error of a Distribution*

Recalling that 50 per cent. of the number of items fall within a range of the lower and upper quartiles, and that the quartile distance in a normal distribution is equal to .6745 of the standard deviation, it will be seen that by multiplying the standard deviation by .6745 we have a quantity which when measured from both sides of the mean includes one-half of the total number of items. The formula, then, for the probable error of a distribution is  $.6745 \sigma$ . This means that in any other distributions of like or similar data the chances are even that one-half of the number of items will fall within a range of the mean of the first distribution plus or minus .6745 of the standard deviation. For instance, if the mean is 5 and the standard deviation is 2 the indications are that, on an average, in any additional like or similar distributions 50 per cent. of the items will lie between 3.65 and 6.35.<sup>1</sup>

<sup>1</sup> For additional formulas for the probable error of other statistical constants the reader is referred to *Handbook of Mathematical Statistics*, 1924, page 77, prepared by members of the Committee on Mathematical Analysis of Statistics of the Division of Physical Sciences of the National Research Council.

## CHAPTER X

### LONG-TIME TRENDS

The general course or tendency of a time series over a long period of years is known as a long-time trend. In fitting a line to a series of this type to designate the trend there are several factors which will determine the method to be followed, important among which are the nature of the data themselves, the length of the period involved, and the purpose for which the calculated trend is intended. If there are cyclical movements, that is, if the data show two or more distinct tendencies to move in the same or opposite directions, the type of line or curve to be fitted will be entirely different than in those instances where there is but one such distinct tendency. When only an approximation of the general movement is desired, less care may be given in fitting the trend than is necessary if exact measurement is to be made. The principal significance of any trend, as indicated, is that it merely shows the general inclination or direction in the movement of a series, so that in many instances the upward or downward tendencies may be graphically presented without any great amount of mathematical calculation. It is only when precise measurement is to be made of the degree or extent of the rise and fall in the course of a series that rigid formulas are necessary. A few of the common methods of fitting trends in ordinary statistical analysis will be discussed and illustrated to give the student a conception of the technique involved.

## FREE-HAND TREND

Fitting of a trend by the free-hand method consists merely of drawing a line through the data in such a way that it appears to show the best approximation of the general course of the series. In Table LXI are shown the statistics of yield of tobacco per acre in the United States for a period of 19 years, and in Chart 29 these yields are presented graphically. For a general idea as to the trend of these yields over the nineteen-year period a line may be drawn from the extreme left to the right in such a way that it represents, as best the eye can detect, the course of the series. Such a free-hand trend is represented by the dotted line in Chart 29.

TABLE LXI. AVERAGE YIELD OF TOBACCO PER ACRE IN THE UNITED STATES FOR NINETEEN YEARS, 1910-1928 <sup>1</sup>

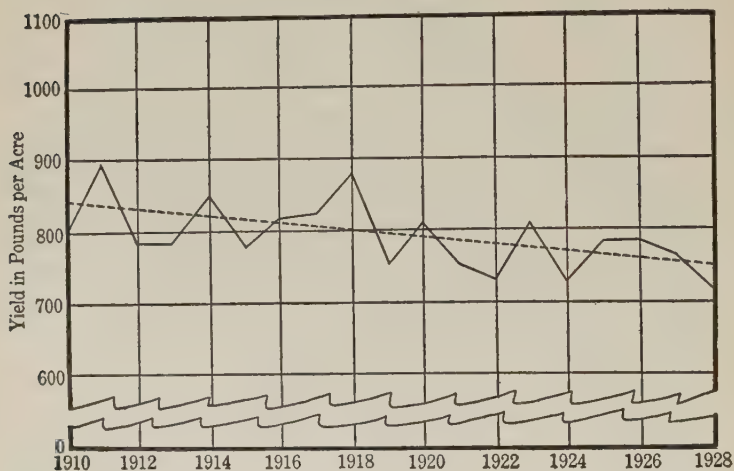
| YEAR | AVERAGE YIELD PER ACRE |
|------|------------------------|
|      | <i>Pounds</i>          |
| 1910 | 808                    |
| 1911 | 894                    |
| 1912 | 786                    |
| 1913 | 784                    |
| 1914 | 846                    |
| 1915 | 775                    |
| 1916 | 816                    |
| 1917 | 823                    |
| 1918 | 874                    |
| 1919 | 751                    |
| 1920 | 807                    |
| 1921 | 750                    |
| 1922 | 736                    |
| 1923 | 807                    |
| 1924 | 734                    |
| 1925 | 783                    |
| 1926 | 784                    |
| 1927 | 765                    |
| 1928 | 718                    |

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 893, Table 341.

CHART 29. AVERAGE YIELD OF TOBACCO PER ACRE IN THE  
UNITED STATES FOR NINETEEN YEARS, 1910-1928

(Based on data in Table LXI)

(FREE-HAND TREND)



### MOVING AVERAGE

The moving average, sometimes called the running average, is merely a number of averages calculated from a series by the simple process of dropping an item every time one is added. The important feature regarding its calculation is that, so far as practicably possible, the number of items of which the average is calculated should coincide with the cycle, if there is one. To illustrate, if we have a series of fifteen items with the first five showing a general trend upward, the next five showing a trend downward, and the third group of five showing a trend upward, it would be logical to base the moving average on five items.

In Table LXII a three-year moving average is calculated for tobacco yields in the United States for nineteen

years, and it is presented graphically in Chart 30, being represented by the dotted line. If the average had been calculated on the basis of five-year periods the first ordinate would have been plotted for 1912 instead of for 1911, and for the midpoint between 1911 and 1912 if it had been a four-year moving average. The essential feature to remember is that the first ordinate of the moving trend is always plotted for the midyear of the series for which the average is calculated, or for any other midpoint that may be represented on the x axis.

TABLE LXII. THREE-YEAR MOVING AVERAGE OF YIELD OF TOBACCO PER ACRE IN THE UNITED STATES, 1910-1928 <sup>1</sup>

| 1    | 2                           | 3                   | 4                            |
|------|-----------------------------|---------------------|------------------------------|
| YEAR | YIELD PER ACRE<br>IN POUNDS | THREE-YEAR<br>TOTAL | THREE-YEAR MOVING<br>AVERAGE |
| 1910 | 808                         |                     |                              |
| 1911 | 894                         | 2,488               | 829                          |
| 1912 | 786                         | 2,464               | 821                          |
| 1913 | 784                         | 2,416               | 805                          |
| 1914 | 846                         | 2,405               | 802                          |
| 1915 | 775                         | 2,437               | 812                          |
| 1916 | 816                         | 2,414               | 805                          |
| 1917 | 823                         | 2,513               | 838                          |
| 1918 | 874                         | 2,448               | 816                          |
| 1919 | 751                         | 2,432               | 811                          |
| 1920 | 807                         | 2,308               | 769                          |
| 1921 | 750                         | 2,293               | 764                          |
| 1922 | 736                         | 2,293               | 764                          |
| 1923 | 807                         | 2,277               | 759                          |
| 1924 | 734                         | 2,324               | 775                          |
| 1925 | 783                         | 2,301               | 767                          |
| 1926 | 784                         | 2,332               | 777                          |
| 1927 | 765                         | 2,267               | 756                          |
| 1928 | 718                         |                     |                              |

If it is desired to have an ordinate for 1910 and 1928 an average may be taken of the first and last two items in the

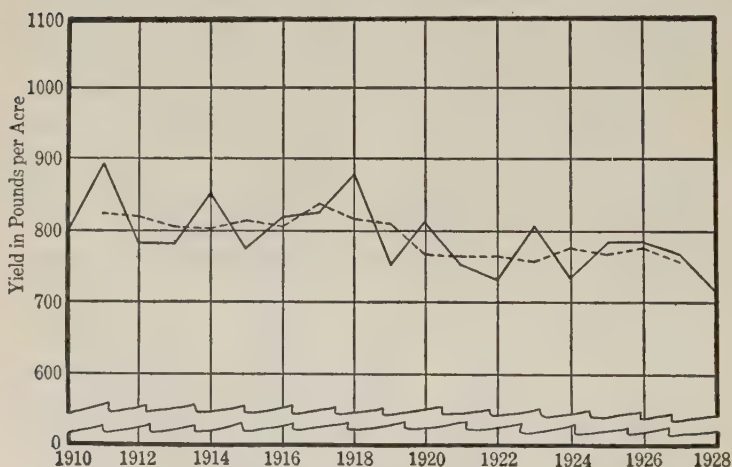
<sup>1</sup> Based on original data in Table LXI, page 163.

series, respectively. Thus, for 1910 the ordinate would be 851 and for 1928 it would be 767. These kinds of calculations, however, may not be valid because the averages are based on fewer items than are the others of the series.

CHART 30. THREE-YEAR MOVING AVERAGE YIELD OF TOBACCO PER ACRE  
IN THE UNITED STATES, 1910-1928

(Based on data in Table LXII)

(MOVING AVERAGE)



Although of no great value when exact numerical measurement is to be made of the trend in a series, the moving average does have a decided significance in certain instances. It is especially useful in making approximations of the general movements in a series, and particularly in eliminating a large part of a cycle that is rather regular. The fitting of free-hand trends is also often facilitated by the application of the moving average as a preliminary step. The principal objectionable features of the use of this average in presenting trends lie in the fact that it is



not very satisfactory when there is no pronounced cycle, and when the curve representing the actual values in the series shows sudden or marked changes in any direction that are not a part of the general cyclical movement or tendency.

#### SEMI-AVERAGE

By this method of approximating the general trend of a series, the group is divided into two equal parts and an average of each one calculated, the averages located on the y axis, and a straight line drawn through them and extending on either side to the extremities of the data. Let us assume we have a series of ten items plotted similar to Chart 30. Dividing these items into two equal parts gives us five in each group. Then each group is summated and the totals divided by 5. The averages thus obtained are to be plotted as ordinates on the y axis at the mid-points of the respective groups. Thus, the average of the first five items will become the y ordinate for the third value on the x axis and the average of the second group will be the y ordinate of the eighth x value.

Let us further illustrate the semi-average method of fitting a trend by turning back to Table LXII. Here are nineteen items. A division into two groups may place the first nine in one group, which includes the yields from 1910 to 1918, inclusive, and the last nine, including the yields from 1920 to 1928, in another. By this grouping the yield for 1919 is completely ignored. Averages calculated for these two groups become the y ordinates for the years 1914 and 1924, respectively. Another method of grouping would be to add the yield for 1919 with both groups, thus giving ten items in each one. In this instance the average for the ten-year period from 1910 to 1919 would become the y ordinate for the midpoint on the x axis between 1914 and

1915. The average for the second ten-year period, 1919 to 1928, would become the y ordinate for the midpoint on the x axis lying between 1923 and 1924. There is still a third grouping that is possible. The yield for 1919 may be divided by 2, so that the total of 19 items will be separated into two groups of  $9\frac{1}{2}$  items each. The averages calculated on the basis of this grouping would, in the first instance, become the y ordinate for the point on the x axis lying at three-fourths of the distance between 1913 and 1914, and for the second group the average would be the y ordinate for the point on the x axis lying three-fourths of the distance between 1923 and 1924. There is not likely to be much difference in the ordinates of the trend as fitted by these several methods. Perhaps it will be found easier where there is an odd number of items in a series to ignore the middle item in calculating the averages of the two groups. Thus, in a series of nine items the fifth one may be ignored, the sixth one may be ignored in a series of eleven items, and so on.

Since by the semi-average method of fitting a trend the items in a series are divided into two equal parts, it is satisfactory only when there is a consistent tendency in the direction of movement in the series. As a matter of fact, it is, at the best, only an approximation, and it must be applied with caution, a principal objectionable feature in its use in making this approximation being that either of the ordinates of the line may be greatly affected by a single unusually small or large item.

#### STRAIGHT LINE TREND OF LEAST SQUARES

The straight line of least squares, sometimes called the first degree parabola, is so fitted that the summation of the squares of the deviations from it is less than the sum-

mation of the squares of deviations from any other straight line that can be drawn. It is this characteristic, therefore, from which is derived the name of "least squares." This line, of course, can properly be fitted only to those series which over a long period show a general movement in one direction. It could not, for example, be applied to a series which first shows a marked tendency upward and then moves downward at a decided slope.

There are two common methods by which the straight line of least squares may be calculated. The first of these methods is illustrated in Table LXIII. In column 1 are the years. Column 2 shows the data ( $y$ ) to which the trend is to be fitted. The midpoint of column 2 falls on the year 1919. The deviations from this midpoint are recorded in column 3. Column 4 shows the products of corresponding pairs of items in columns 2 and 3. In column 5 are the squares of the deviations in column 3, and in column 6 are the ordinates of the straight line trend of least squares as fitted to the data in column 2.

The calculations involved in obtaining the values of  $Y$ , the ordinates of trend, in column 6 are very simple. First, we locate the middle item in the series. This is found to be the yield for 1919, which becomes the point of origin. In column 3 the point of origin is designated as 0. Each year preceding 1919 is a minus deviation, and so we have the minus deviations from 1 to 9. Each year following 1919 is a plus deviation, as shown in the lower half of column 3. The next step is to multiply the items in the series by these deviations. Accordingly, the yields per acre in column 2 are multiplied by the corresponding deviations and the products recorded in column 4. As will be observed, these products carry the same sign as the deviations in column 3. They are, therefore, summated alge-

braically, the minus quantities in one column and the plus quantities in another. The smaller total is then subtracted from the larger and the result recorded as the summation of  $xy$ , which is the summation of column 4. This column shows a total of  $-2,747$ . It is this total which when divided by the summation of the deviations squared gives the slope of the line. Proceeding, therefore, we divide  $-2,747$  by  $570$ , the summation of  $x^2$ , and obtain a value of approximately  $-5$ .<sup>1</sup> This slope means that on an average the yield of tobacco declined at the rate of five pounds a year from 1910 to 1928. It is this quantity, then, that represents the slope of the straight line of least squares.

We have already seen that the point of origin on the line coincides with the year 1919. The ordinate of the line for that year is the mean of the data  $y$  in column 2. Summating column 2 and dividing by 19, the number of items, gives a mean of 792. This, as will be noted in column 6, is recorded as the value of  $Y$  for 1919. Recalling that the slope of the line is  $-5$ , we calculate the other ordinates by the formula  $Y = a + bx$ , in which  $Y$  is the ordinate of the line,  $a$  the arithmetic average,  $b$  the slope of the line, and  $x$  the deviation from the point of origin. Let us illustrate the calculation of these other ordinates by referring to the year 1918. Column 3 for that year shows a deviation of  $-1$ . Therefore, we multiply  $-5$ , the slope, by  $-1$  and add the product to the mean. Since both deviation and slope carry the minus sign the product obtained by multiplication is a plus. Thus we have 5 as the product of  $-1$  and  $-5$ , which added to 792 gives 797 as the ordinate for 1918. The other ordinates of the line are calculated in the same way, the deviations in column 3 increasing by 1 each time. The ordinates of the line

<sup>1</sup> In some instances it may be advisable to carry fractions.

above the point of origin, including the years 1920 to 1928, will decrease as we proceed from 1919 because, as is seen, the deviations in column 3 carry the plus sign, while the slope is a minus quantity.

Instead of multiplying the slope by the deviation corresponding to each year we may start with the mean as the ordinate for 1919, add 5 to obtain 797 for 1918, add 5 to obtain 802 for 1917, add 5 to 802 to obtain 807 for 1916, and so on. Then for the ordinates above 1919 we may begin by subtracting 5 from 792 to obtain 787 as the ordinate for 1920, subtract 5 from 787 to obtain 782 for 1921, subtract 5 from 782 to obtain 777 for 1922, and so on for the other ordinates. It may be mentioned also that instead of calculating each ordinate both below and above the mean we may determine one on either side, plot them, and then draw a straight line passing through the mean and connecting the two ordinates as calculated. For example, we may determine the ordinate of trend for 1910 by multiplying the slope of the line,  $-5$ , by  $-9$ , the deviation in column 3 for 1910, and then add the product to the mean for the ordinate of that year. Similarly, the ordinate for 1928 may be calculated by subtracting 45 from 792. A line connecting these two ordinates will pass through the mean. In calculating only two ordinates, one on either side of the mean, they should be preferably the ones corresponding to the first and last  $x$  values of the series.

The student will find as the method becomes well understood that with a calculating machine the plus and minus summations of column 4 can be obtained without taking each individual product from the machine. In like manner, the summation of column 5 may be determined by squaring each deviation and carrying the products to the end.



TABLE LXIII. STRAIGHT LINE TREND OF YIELD OF TOBACCO PER ACRE  
IN THE UNITED STATES, 1910-1928<sup>1</sup>

(Line of Least Squares)

| 1     | 2                                                 | 3                                 | 4                                                                                                     | 5                                                                          | 6                                                                    |
|-------|---------------------------------------------------|-----------------------------------|-------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------|
| YEAR  | YIELD PER<br>ACRE IN<br>POUNDS<br>(Mean =<br>792) | DEVIATION<br>FROM THE<br>MIDPOINT | PRODUCT OF<br>YIELD PER ACRE<br>AND DEVIATION<br>FROM THE<br>MIDPOINT<br>(Column 2 Times<br>Column 3) | SQUARE OF<br>DEVIATIONS<br>FROM THE<br>MIDPOINT<br>(Square of<br>Column 3) | ORDINATES<br>OF THE<br>STRAIGHT<br>LINE TREND<br>OF LEAST<br>SQUARES |
|       | y                                                 | x                                 | xy                                                                                                    | x <sup>2</sup>                                                             | Y                                                                    |
| 1910  | 808                                               | -9                                | -7,272                                                                                                | 81                                                                         | 837                                                                  |
| 1911  | 894                                               | -8                                | -7,152                                                                                                | 64                                                                         | 832                                                                  |
| 1912  | 786                                               | -7                                | -5,502                                                                                                | 49                                                                         | 827                                                                  |
| 1913  | 784                                               | -6                                | -4,704                                                                                                | 36                                                                         | 822                                                                  |
| 1914  | 846                                               | -5                                | -4,230                                                                                                | 25                                                                         | 817                                                                  |
| 1915  | 775                                               | -4                                | -3,100                                                                                                | 16                                                                         | 812                                                                  |
| 1916  | 816                                               | -3                                | -2,448                                                                                                | 9                                                                          | 807                                                                  |
| 1917  | 823                                               | -2                                | -1,646                                                                                                | 4                                                                          | 802                                                                  |
| 1918  | 874                                               | -1                                | - 874                                                                                                 | 1                                                                          | 797                                                                  |
| 1919  | 751                                               | 0                                 | 0                                                                                                     | 0                                                                          | 792                                                                  |
| 1920  | 807                                               | 1                                 | 807                                                                                                   | 1                                                                          | 787                                                                  |
| 1921  | 750                                               | 2                                 | 1,500                                                                                                 | 4                                                                          | 782                                                                  |
| 1922  | 736                                               | 3                                 | 2,208                                                                                                 | 9                                                                          | 777                                                                  |
| 1923  | 807                                               | 4                                 | 3,228                                                                                                 | 16                                                                         | 772                                                                  |
| 1924  | 734                                               | 5                                 | 3,670                                                                                                 | 25                                                                         | 767                                                                  |
| 1925  | 783                                               | 6                                 | 4,698                                                                                                 | 36                                                                         | 762                                                                  |
| 1926  | 784                                               | 7                                 | 5,488                                                                                                 | 49                                                                         | 757                                                                  |
| 1927  | 765                                               | 8                                 | 6,120                                                                                                 | 64                                                                         | 752                                                                  |
| 1928  | 718                                               | 9                                 | 6,462                                                                                                 | 81                                                                         | 747                                                                  |
| Total | —                                                 | 0                                 | -2,747                                                                                                | 570                                                                        | —                                                                    |

Slope of the straight line trend of least squares

$$(b) = \frac{\Sigma xy}{\Sigma x^2} = \frac{-2,747}{570} = -5.$$

Ordinates of the straight line trend of least squares

$$(Y) = a + bx.$$

The data in column 2 of Table LXIII are shown graphically by the continuous line in Chart 31, and the

<sup>1</sup> Based on original data in Table LXI, page 163.



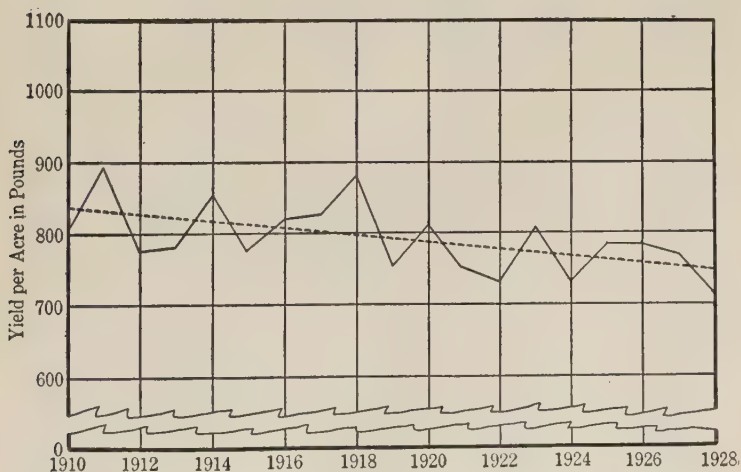
straight line trend of least squares, plotted on the ordinates of column 6, is represented by the dotted line.

Before passing from this phase of our subject, attention must be called to a modification of the method illustrated in Table LXIII that is necessary when there is an even number of items in the series to which the straight line is being fitted. Perhaps this can best be done by showing the actual calculations involved. For this purpose let us observe the technique of Table LXIV.

CHART 31. STRAIGHT LINE TREND OF THE YIELD OF TOBACCO PER ACRE  
IN THE UNITED STATES, 1910-1928

(Based on data in Table LXIII)

(LINE OF LEAST SQUARES)



Slope of the straight line trend of least squares,  
Table LXIV,  $(b) = \frac{\sum xy}{\sum x^2} = \frac{22.50}{82.50} = .27$

Ordinates of the straight line trend of least squares,  
Table LXIV,  $(Y) = a + bx.$

TABLE LXIV. STRAIGHT LINE TREND OF THE YIELD OF WHEAT PER ACRE IN THE UNITED STATES FOR TEN YEARS, 1919-1928<sup>1</sup>

| 1     | 2                                                | 3                                  | 4                                                                                                           | 5                                                                          | 6                                                                    |
|-------|--------------------------------------------------|------------------------------------|-------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------|
| YEAR  | YIELD PER<br>ACRE IN<br>BUSHEL<br>(Mean = 14.12) | DEVIATIONS<br>FROM THE<br>MIDPOINT | PRODUCT OF<br>YIELD PER<br>ACRE AND<br>DEVIATION<br>FROM THE<br>MIDPOINT<br>(Column 2<br>Times<br>Column 3) | SQUARE OF<br>DEVIATIONS<br>FROM THE<br>MIDPOINT<br>(Square of<br>Column 3) | ORDINATES<br>OF THE<br>STRAIGHT LINE<br>TREND OF<br>LEAST<br>SQUARES |
|       | y                                                | x                                  | xy                                                                                                          | x <sup>2</sup>                                                             | Y                                                                    |
| 1919  | 12.8                                             | -4.5                               | -57.60                                                                                                      | 20.25                                                                      | 12.91                                                                |
| 1920  | 13.6                                             | -3.5                               | -47.60                                                                                                      | 12.25                                                                      | 13.18                                                                |
| 1921  | 12.8                                             | -2.5                               | -32.00                                                                                                      | 6.25                                                                       | 13.45                                                                |
| 1922  | 13.9                                             | -1.5                               | -20.85                                                                                                      | 2.25                                                                       | 13.72                                                                |
| 1923  | 13.4                                             | -.5                                | -6.70                                                                                                       | .25                                                                        | 13.99                                                                |
| 1924  | 16.5                                             | .5                                 | 8.25                                                                                                        | .25                                                                        | 14.26                                                                |
| 1925  | 12.9                                             | 1.5                                | 19.35                                                                                                       | 2.25                                                                       | 14.53                                                                |
| 1926  | 14.8                                             | 2.5                                | 37.00                                                                                                       | 6.25                                                                       | 14.80                                                                |
| 1927  | 14.9                                             | 3.5                                | 52.15                                                                                                       | 12.25                                                                      | 15.07                                                                |
| 1928  | 15.6                                             | 4.5                                | 70.20                                                                                                       | 20.25                                                                      | 15.34                                                                |
| Total | —                                                | 0                                  | 22.20                                                                                                       | 82.50                                                                      | —                                                                    |

We see, therefore, since the slope of the line is a plus quantity, the yield of wheat has shown an upward tendency since 1919. It will be recalled that in determining the ordinates of the straight line the mean is plotted as the ordinate of the midpoint of the series. The mean of column 2 in Table LXIV is 14.12, and the midpoint of the series falls midway between 1923 and 1924. The ordinate for this point, then, is 14.12. To move from it to 1923 is but a deviation of  $-.5$ , to 1922 is a deviation of  $-1.5$ , and so on to 1919. Proceeding beyond the mean, or median ordinate on the line, it is seen that from the midpoint of 1923 and 1924 to 1924 is a deviation of  $.5$ , to 1925 the deviation is  $1.5$ , with the deviation increasing by 1 for each additional year.

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 670, Table 1.

Determination of the ordinates is done by substituting the proper values in the formula  $Y = a + bx$ , in which, as we have seen already,  $Y$  is the ordinate,  $a$  the arithmetic mean of the series,  $b$  the slope of the line of least squares, and  $x$  the deviation from the point of origin.

If the yearly trend is calculated for data made up of monthly items, such as yearly prices of wheat and corn, the monthly slope may be approximated from the yearly slope by dividing by 12, the number of months in a year.

TABLE LXV. AVERAGE YIELD OF TOBACCO PER ACRE IN THE UNITED STATES, 1910-1928 <sup>1</sup>

| 1     | 2                           | 3                                        | 4                                                                                    | 5                                                                                                       |
|-------|-----------------------------|------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| YEAR  | YIELD PER ACRE<br>IN POUNDS | DEVIATION FROM<br>THE POINT OF<br>ORIGIN | SQUARE OF THE<br>DEVIATIONS FROM<br>THE POINT OF OR-<br>IGIN (Square of<br>Column 3) | PRODUCT OF DE-<br>VIATIONS FROM THE<br>POINT OF ORIGIN<br>AND YIELD (Col-<br>umn 3 Times Col-<br>umn 2) |
|       | y                           | x                                        | x <sup>2</sup>                                                                       | xy                                                                                                      |
| 1910  | 808                         | 1                                        | 1                                                                                    | 808                                                                                                     |
| 1911  | 894                         | 2                                        | 4                                                                                    | 1,788                                                                                                   |
| 1912  | 786                         | 3                                        | 9                                                                                    | 2,358                                                                                                   |
| 1913  | 784                         | 4                                        | 16                                                                                   | 3,136                                                                                                   |
| 1914  | 846                         | 5                                        | 25                                                                                   | 4,230                                                                                                   |
| 1915  | 775                         | 6                                        | 36                                                                                   | 4,650                                                                                                   |
| 1916  | 816                         | 7                                        | 49                                                                                   | 5,712                                                                                                   |
| 1917  | 823                         | 8                                        | 64                                                                                   | 6,584                                                                                                   |
| 1918  | 874                         | 9                                        | 81                                                                                   | 7,866                                                                                                   |
| 1919  | 751                         | 10                                       | 100                                                                                  | 7,510                                                                                                   |
| 1920  | 807                         | 11                                       | 121                                                                                  | 8,877                                                                                                   |
| 1921  | 750                         | 12                                       | 144                                                                                  | 9,000                                                                                                   |
| 1922  | 736                         | 13                                       | 169                                                                                  | 9,568                                                                                                   |
| 1923  | 807                         | 14                                       | 196                                                                                  | 11,298                                                                                                  |
| 1924  | 734                         | 15                                       | 225                                                                                  | 11,010                                                                                                  |
| 1925  | 783                         | 16                                       | 256                                                                                  | 12,528                                                                                                  |
| 1926  | 784                         | 17                                       | 289                                                                                  | 13,328                                                                                                  |
| 1927  | 765                         | 18                                       | 324                                                                                  | 13,770                                                                                                  |
| 1928  | 718                         | 19                                       | 361                                                                                  | 13,642                                                                                                  |
| Total | 15,041                      | 190                                      | 2,470                                                                                | 147,663                                                                                                 |

<sup>1</sup> Data from Table LXI, page 163.

The second method by which the straight line trend may be calculated consists of solving the two normal equations of the formula  $Y = a + bx$ , in which  $Y$  is the ordinate of the straight line,  $a$  the starting point,  $b$  the slope of the line, and  $x$  the distance from the point of origin. Let us observe the calculation in Table LXV.

For the formula  $Y = a + bx$  the following normal equations may be constructed:

$$\begin{aligned}\Sigma y &= an + b \Sigma x \\ \Sigma xy &= a \Sigma x + b \Sigma x^2.\end{aligned}$$

Substituting the proper values and solving the two normal equations we have:

1.  $19a + 190b = 15,041$
2.  $190a + 2470b = 147,663$
3.  $190a + 1900b = 150,410$  (Equation 1 times 10)
4.  $190a + 2470b = 147,663$  (Equation 2 times 1)
5.  $\quad \quad - 570b = 2,747$  (Equation 3 minus equa-
6.  $\quad \quad \quad b = -5$       tion 4)
7.  $19a + 190b = 15,041$
8.  $19a + 190(-5) = 15,041$
9.  $19a - 950 = 15,041$
10.  $19a = 15,991$  ( $15,041 + 950$ )
11.  $a = 842.$

As will be seen, the value of  $b$  is  $-5$  and the value of  $a$  is 842. Recalling that  $b$  is the slope of the line and that  $a$  is the starting point, all that need be done in determining the ordinates of the straight line of least squares is to substitute actual values in the formula  $Y = a + bx$ . Therefore, to obtain the ordinate for 1910 in Table LXV we multiply the value of  $b$ , which is  $-5$ , by  $x$  and add to the value of  $a$ , which is 842. Proceeding,  $-5$  times 1 equals  $-5$ , which

when added algebraically to 842 gives 837, the ordinate of the line for 1910. Turning to Table LXIII we find the same ordinate was obtained by the first method illustrated. The ordinates for the years following are obtained in the same way as the ordinate for 1910, the value of  $x$  increasing by 1 for each year above 1910.

Unless the student can readily solve normal equations the method of fitting a straight line as outlined in Tables LXIII and LXIV will be found more desirable. It may be pointed out that by the method outlined in these tables  $\Sigma x$ , column 3, is zero, and that  $a\Sigma x$  and  $b\Sigma x$  disappear, with the result that  $an = \Sigma y$  and  $b\Sigma x^2 = \Sigma xy$ . When  $\Sigma y$  is divided by  $n$  the value for  $a$  is found as the midpoint of the line or the arithmetic average. Also, when  $\Sigma x^2$  is divided into  $\Sigma xy$  the value of  $b$ , or the slope of the line, is solved for.

Where the straight line trend of least squares is judiciously applied it is very satisfactory for presenting graphically the general tendency in the direction of a series, but its use can be easily abused, thus invalidating any conclusions that may be drawn from it. This is particularly true when there are marked changes in individual items in a series that otherwise shows a consistent movement in a certain direction. Thus, for example, in superimposing a straight line of least squares upon a chart in which is plotted the curve of original values the slope of the straight line may be greatly affected or modified if a single item shows a marked deviation from items immediately preceding or immediately following, or even a marked deviation from the general direction of trend. This objectionable feature is all the more appreciated when it is recalled that the dividend used in calculating the slope of the line, as illustrated in Table LXIII, is derived from the sum-

mation of the products of the original items and the deviations from the point of origin on the line.

### LOGARITHMIC PARABOLA

When the movement or course of a series is such that it cannot be properly represented by a straight line a trend or line of some other order must be calculated, the type depending upon the nature of the data. One of the most common of these for many series of data is the logarithmic parabola, the formula for which is

$$a+bx+c \log. x = Y.^1$$

The values for substitution in this formula are calculated in Table LXVI.

TABLE LXVI. NUMBER OF FARMERS OF A TOTAL OF 189 INTERVIEWED IN EASTERN NORTH CAROLINA WHO PRODUCED POULTRY FOR OTHER THAN DOMESTIC PURPOSES, 1917-1926 <sup>2</sup>

(Values for Logarithmic Equation)

| 1     | 2                              | 3  | 4      | 5              | 6        | 7                   | 8        | 9     |
|-------|--------------------------------|----|--------|----------------|----------|---------------------|----------|-------|
| YEAR  | y<br>(Number<br>of<br>Farmers) | x  | Log. x | x <sup>2</sup> | y log. x | Log. x <sup>2</sup> | x log. x | xy    |
| 1917  | 59                             | 1  | .0000  | 1              | 0        | .000                | —        | 59    |
| 1918  | 91                             | 2  | .3010  | 4              | 27.4     | .091                | .6020    | 182   |
| 1919  | 108                            | 3  | .4771  | 9              | 51.5     | .228                | 1.4313   | 324   |
| 1920  | 123                            | 4  | .6021  | 16             | 74.1     | .363                | 2.4084   | 492   |
| 1921  | 135                            | 5  | .6990  | 25             | 94.4     | .489                | 3.4950   | 675   |
| 1922  | 144                            | 6  | .7782  | 36             | 112.1    | .606                | 4.6692   | 864   |
| 1923  | 153                            | 7  | .8451  | 49             | 129.3    | .714                | 5.9157   | 1,071 |
| 1924  | 159                            | 8  | .9031  | 64             | 143.6    | .816                | 7.2248   | 1,272 |
| 1925  | 161                            | 9  | .9542  | 81             | 153.6    | .910                | 8.5878   | 1,449 |
| 1926  | 170                            | 10 | 1.0000 | 100            | 170.0    | 1.000               | 10.0000  | 1,700 |
| Total | 1,303                          | 55 | 6.5598 | 385            | 956.0    | 5.217               | 44.3342  | 8,088 |

<sup>1</sup> Sometimes referred to as the second degree logarithmic parabola.

<sup>2</sup> Specially selected data from reports of a survey made in nine counties of North Carolina.



For the logarithmic formula  $a+bx+c \log. x=Y$  the following three equations may be constructed:

1.  $an+b\Sigma x+c\Sigma \log. x = \Sigma y$
2.  $a\Sigma x+b\Sigma x^2+c\Sigma x \log. x = \Sigma xy$
3.  $a\Sigma \log. x+b\Sigma x \log. x+c\Sigma (\log. x^2) = \Sigma y \log. x$

Substituting the proper values as calculated in Table LXVI we have the following equations.

1.  $10a+55b+6.5598c = 1,303$
  2.  $55a+385b+44.3342c = 8,088$
  3.  $6.5598a+44.3342b+5.217c = 956$
- 

Solving for the value of  $c$ :

4.  $55a+302.5b+36.0789c=7,166.5$  (Equation 1 times 5.5)
5.  $55a + 385b+44.3342c=8,088$  (Equation 2 times 1)

---

6.  $82.5b+8.2553c=921.5$  (Equation 5 minus equation 4)
7.  $55a+385b+44.3342c = 8,088$  (Equation 2 times 1)
8.  $55a+372b+43.7414c = 8,015$  (Equation 3 times 8.3844)

---

9.  $13b + .5928c = 73$  (Equation 7 minus equation 8)
10.  $69.5b+7.6625c = 848.5$  (Equation 6 minus equation 9)
11.  $69.5b+6.9545c = 776.3$  (Equation 6 times .842424)

---

12.  $.7080c = 72.2$  (Equation 10 minus equation 11)

$$13. \quad c = 102 \quad (72 \text{ divided by } .7080)$$


---

Solving for the value of  $b$ :

$$14. \quad 82.5b + 8.2553c = 921.5 \text{ (Same as equation 6)}$$

$$15. \quad 82.5b + (8.2553 \times 102) = 924.54$$

$$16. \quad 924.54b = 921.5$$

$$17. \quad b = .9967 \text{ (921.5 divided by 924.54)}$$


---

Solving for the value of  $a$ :

$$18. \quad 55a + 385b + 44.3342c = 8,088 \text{ (Same as equation 2)}$$

$$19. \quad 55a + (385 \times .9967) + (44.3342 \times 102) = 8,088$$

$$20. \quad 55a = 3,182$$

$$21. \quad a = 57.85 \text{ (3,182 divided by 5.5)}$$


---

With the values of  $a$ ,  $b$ , and  $c$  calculated, the ordinates of the second degree logarithmic trend may now be determined by substitution in the formula  $Y = a + bx + c \log x$ . The technique involved consists merely of adding to the value of  $a$ , 57.85, the summation of the products obtained by multiplying the value of  $b$ , .9967, by  $x$  (column 3 in Table LXVI) and the value of  $c$ , 102, by the logarithm of  $x$  as given in column 4 of the table.

$$\text{Ordinate for 1917} = 57.85 + (.9967 \times 1) + (102 \times 0) = 58.847$$

$$\text{Ordinate for 1918} = 57.85 + (.9967 \times 2) + (102 \times .3010) = 90.545$$

$$\text{Ordinate for 1919} = 57.85 + (.9967 \times 3) + (102 \times .4771) = 109.504$$

$$\text{Ordinate for 1920} = 57.85 + (.9967 \times 4) + (102 \times .6021) = 123.251$$

$$\text{Ordinate for 1921} = 57.85 + (.9967 \times 5) + (102 \times .6990) = 134.132$$

$$\text{Ordinate for 1922} = 57.85 + (.9967 \times 6) + (102 \times .7782) = 143.207$$

$$\text{Ordinate for 1923} = 57.85 + (.9967 \times 7) + (102 \times .8451) = 151.027$$

$$\text{Ordinate for 1924} = 57.85 + (.9967 \times 8) + (102 \times .9031) = 157.940$$

$$\text{Ordinate for 1925} = 57.85 + (.9967 \times 9) + (102 \times .9542) = 164.149$$

$$\text{Ordinate for 1926} = 57.85 + (.9967 \times 10) + (102 \times 1.0000) = 169.817$$

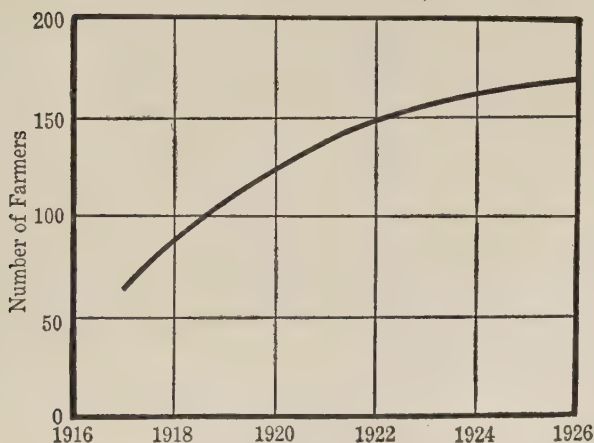
By comparing these ordinates with the data in column 2 of Table LXVI it will be seen that the ordinates of the calculated logarithmic trend deviate only slightly from the actual values. We conclude, therefore, that the loga-

CHART 32. LOGARITHMIC TREND OF NUMBER OF FARMERS IN A TOTAL OF 189 INTERVIEWED WHO PRODUCED POULTRY FOR OTHER THAN DOMESTIC PURPOSES, 1917-1926

(Based on data in Table LXVI)

$$(Y = a + bx + c \log.x)$$

(LOGARITHMIC PARABOLA)



arithmic parabola is a much better fit to the data than the straight line of least squares would be. Chart 32 shows the logarithmic trend of the data in column 2 of Table LXVI. The actual values are not plotted because they so nearly coincide with the ordinates of the calculated trend.

### SECOND DEGREE PARABOLA

The ordinates of the ordinary second degree parabola may be determined by the formula  $Y = a + bx + cx^2$ .

Let us take for illustrative purposes the same data as presented in Table LXVI. The necessary quantities or summations for substitution in arriving at the values of  $a$ ,  $b$ , and  $c$  are calculated as follows:

TABLE LXVII. ORDINATES OF SECOND DEGREE PARABOLIC TREND <sup>1</sup>

| 1     | 2     | 3  | 4     | 5              | 6              | 7                | 8              | 9                |
|-------|-------|----|-------|----------------|----------------|------------------|----------------|------------------|
| YEAR  | y     | x  | xy    | x <sup>2</sup> | x <sup>3</sup> | x <sup>2</sup> y | x <sup>4</sup> | Y<br>(Ordinates) |
| 1917  | 59    | 1  | 59    | 1              | 1              | 59               | 1              | 64.9             |
| 1918  | 91    | 2  | 182   | 4              | 8              | 364              | 16             | 86.2             |
| 1919  | 108   | 3  | 324   | 9              | 27             | 972              | 81             | 104.9            |
| 1920  | 123   | 4  | 492   | 16             | 64             | 1,968            | 256            | 121.1            |
| 1921  | 135   | 5  | 675   | 25             | 125            | 3,375            | 625            | 134.8            |
| 1922  | 144   | 6  | 864   | 36             | 216            | 5,184            | 1,296          | 146.0            |
| 1923  | 153   | 7  | 1,071 | 49             | 343            | 7,497            | 2,401          | 154.6            |
| 1924  | 159   | 8  | 1,272 | 64             | 512            | 10,176           | 4,096          | 160.7            |
| 1925  | 161   | 9  | 1,449 | 81             | 729            | 13,041           | 6,561          | 164.3            |
| 1926  | 170   | 10 | 1,700 | 100            | 1,000          | 17,000           | 10,000         | 165.4            |
| Total | 1,303 | 55 | 8,088 | 385            | 3,025          | 59,636           | 25,333         | —                |

In determining the values of  $a$ ,  $b$ , and  $c$  the following equations are constructed:

1.  $\Sigma y = na + b \Sigma(x) + c \Sigma(x^2)$
2.  $\Sigma xy = a \Sigma(x) + b \Sigma(x^2) + c \Sigma(x^3)$
3.  $\Sigma xy^2 = \Sigma a(x^2) + b \Sigma(x^3) + c \Sigma(x^4)$

All that is necessary now is to make the proper substitutions and solve:

1.  $10a + 55b + 385c = 1,303$
2.  $55a + 385b + 3,025c = 8,088$
3.  $385a + 3,025b + 25,333c = 59,636$

<sup>1</sup> Based on data in Table LXVI, page 178.

$$4. \quad 385a + 2,695b + 21,175c = 56,616 \quad (\text{Equation 2 times 7})$$

$$5. \quad 330b + 4,158c = 3,020 \quad (\text{Equation 3 minus equation 4})$$


---

$$6. \quad 55a + 385b + 3,025c = 8,088 \quad (\text{Equation 2 times 1})$$

$$7. \quad 55a + 302.5b + 2,117.5c = 7,166.5 \quad (\text{Equation 1 times 5.5})$$


---

$$8. \quad 82.5b + 907.5c = 921.5 \quad (\text{Equation 6 minus equation 7})$$

$$9. \quad 330.0b + 3,630.0c = 3,686.0 \quad (\text{Equation 8 times 4})$$

$$10. \quad 330.0b + 4,158.0c = 3,020.0 \quad (\text{Equation 5 times 1})$$


---

$$11. \quad -528.0c = 666.0 \quad (\text{Equation 9 minus equation 10})$$

$$12. \quad c = -1.261$$

Solving for  $b$ :

$$13. \quad 330b + 4,158c = 3,020 \quad (\text{Same as equation 5})$$

$$14. \quad 330b + (4,158 \times -1.261) = 3,020$$

$$15. \quad 330b = 8,263.24$$

$$16. \quad b = 25.04$$

Solving for  $a$ :

$$17. \quad 10a + 55b + 385c = 1,303 \quad (\text{Same as equation 1})$$

$$18. \quad 10a + (55 \times 25.04) + (385 \times -1.261) = 1,303$$

$$19. \quad 10a = 411.29$$

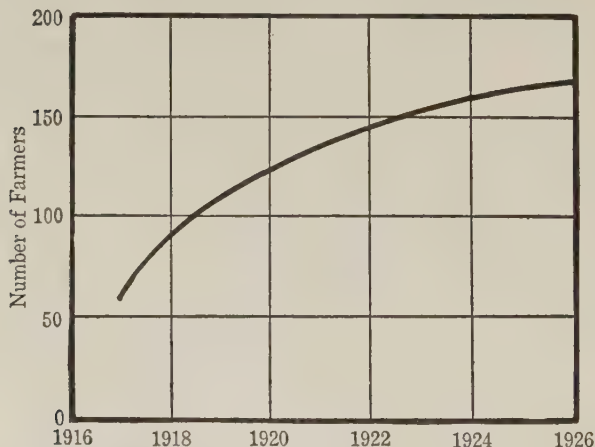
$$20. \quad a = 41.13$$

The ordinates in column 9 of Table LXVII are determined by simply solving  $Y = a + bx + cx^2$ . As we have already calculated the values of  $a$ ,  $b$ , and  $c$ , the solutions are now very simple. All that is necessary is to add to the value of  $a$  the product of the value of  $b$  multiplied by  $x$  and the product of the value of  $c$  multiplied by  $x^2$ . Let us take the year 1919 for illustration. The value of  $a$  we

have already seen to be 41.13. The value of  $b$  is 25.04. Multiplying 25.04 by 3, the  $x$  value for 1919, we obtain 75.12. This product is added to 41.13, giving a total of 116.25. Next, the value of  $c$ ,  $-1.261$ , is multiplied by  $x^2$  for 1919. This gives a product of  $-11.35$ , and since it is a negative quantity it is subtracted from 116.25 to obtain

CHART 33. SECOND DEGREE PARABOLIC TREND OF NUMBER OF FARMERS IN A TOTAL OF 189 INTERVIEWED WHO PRODUCED POULTRY FOR OTHER THAN DOMESTIC PURPOSES, 1917-1926

(Based on ordinates of trend as calculated in Table LXVII)  
( $Y = a + bx + cx^2$ )



104.9 as the ordinate of the line for 1919. All the other ordinates are calculated in the same way. The essential feature to remember is that in the formula  $Y = a + bx + cx^2$  the  $cx^2$  product, obtained by multiplying  $x^2$  by the value of  $c$ , is a minus quantity, which means it is added algebraically. This, of course, is the same as ordinary subtraction. Therefore, in calculating the ordinates we simply add the product of  $bx$  (value of  $b$  times  $x$ ) to the value of  $a$ , and then subtract the product of  $cx^2$  (value of  $c$  times

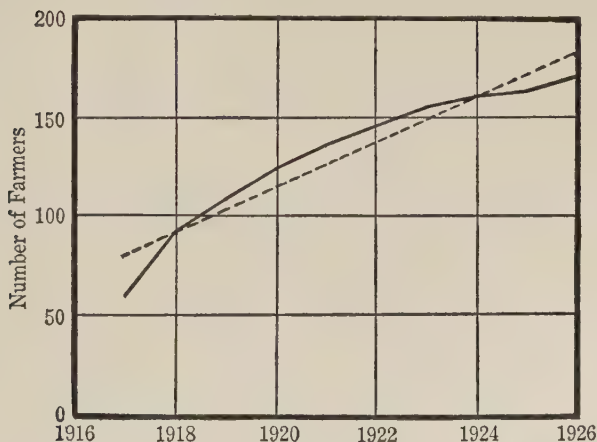


$x^2$ ). If  $c$  were a plus quantity the product obtained in multiplying it by  $x^2$  would be added to 41.13, the value of  $a$ , just the same as the product of  $bx$ .

The more exact nature of the second degree parabola will be observed in Chart 33.

CHART 34. STRAIGHT LINE TREND OF NUMBER OF FARMERS IN A TOTAL OF 189 INTERVIEWED WHO PRODUCED POULTRY FOR OTHER THAN DOMESTIC PURPOSES, 1917-1926

(Based on data in Table LXVI)  
( $Y = a + bx$ )



To show that a straight line does not fit the data in column 2 of Table LXVI as closely as the logarithmic and second degree parabolas, the actual values and the ordinates of least squares are plotted in Chart 34. The dotted line is the trend based on  $Y = a + bx$  (straight line of least squares), while the continuous line represents the data.

#### TRENDS OF HIGHER ORDER

It would be impossible to enter into a discussion of all the formulas by which trends other than those outlined

can be calculated. If it is realized that these curved lines are merely based on  $a+bx$  in some modified form the problem of fitting curves or trends of the higher orders is not so difficult as it may first appear. Just as we have seen that the first degree parabola, the straight line of least squares, is calculated from  $Y=a+bx$ , and that the second degree parabola is based on the expansion of this formula to  $Y=a+bx+cx^2$ , so the third degree parabola is based simply on  $Y=a+bx+cx^2+dx^3$ . Trends of other orders would be based on further expansion of the formula, or upon some modification to suit the particular data.

In Table LXVIII are the indices of wheat yields in Maryland for thirteen years, together with the ordinates of trend as calculated from the formula  $Y=a+bx+cx^2+dx^3$ . The trend and data are graphically presented in Chart 35, the continuous line representing the trend and the dotted line representing the data.

TABLE LXVIII. INDICES OF WHEAT YIELDS IN MARYLAND, 1908-1920 <sup>1</sup>

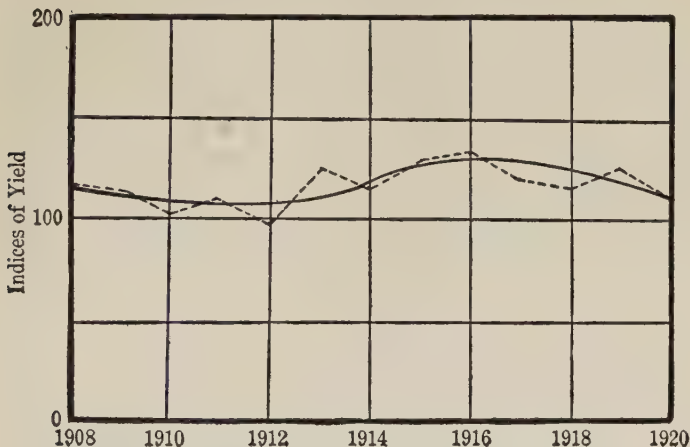
| YEAR | INDICES OF YIELD | ORDINATES OF THIRD DEGREE PARABOLIC TREND |
|------|------------------|-------------------------------------------|
| 1908 | 115              | 113                                       |
| 1909 | 109              | 108                                       |
| 1910 | 101              | 106                                       |
| 1911 | 109              | 106                                       |
| 1912 | 98               | 109                                       |
| 1913 | 125              | 114                                       |
| 1914 | 115              | 118                                       |
| 1915 | 126              | 123                                       |
| 1916 | 134              | 126                                       |
| 1917 | 120              | 127                                       |
| 1918 | 116              | 125                                       |
| 1919 | 125              | 119                                       |
| 1920 | 109              | 109                                       |

<sup>1</sup> Maryland Agricultural Experiment Station records for one variety of wheat.

CHART 35. THIRD DEGREE PARABOLIC TREND OF INDICES OF WHEAT YIELDS IN MARYLAND, 1908-1920

(Based on data in Table LXVIII)

$$(Y = a + bx + cx^2 + dx^3)$$



Obviously, there are so many types of curves that only the more common ones can be explained in detail.<sup>1</sup>

### PRICE DEFLATION

Sometimes in measuring and analyzing the relative trends of series expressed in money values it may be well to take into account the fluctuations in price levels. Let us say, for example, that in 1928 the cost of labor employed in producing an acre of tobacco in the United States was \$80.00, whereas in 1912 the labor cost per acre was \$40.00. Are we to infer from these costs that in 1929 there was two times the quantity of labor employed per acre as in 1912? No, this is not the implication at all. The changes in total labor cost may have been largely due to the increase in wages, rather than to an additional

<sup>1</sup> If the student is interested in the measurement of seasonal variations, interesting discussion and reference will be found in the following sources: *Statistical Methods*, by F. C. Mills, pages 315 and 343, and *Statistical Method*, by Harry Jerome, pages 235 and 246.

amount of actual labor. In 1912 the index number of farm wages was 101,<sup>1</sup> while for 1928 it was 169.<sup>1</sup> It will be seen, therefore, that wages, themselves, do not show an increase of twofold for these years. The logical conclusion, then, is that the amount or number of hours of labor employed in producing an acre of tobacco in 1928 was not twice that of 1912. For comparing the quantity of labor employed in 1928 with that employed in 1912 we deflate the costs by dividing by the index numbers of farm wages. For 1912, therefore, \$40.00 is divided by 101, which gives \$39.61. The cost for 1928, \$80.00, is divided by 169, giving a quotient of \$47.34. These costs are deflated. That is to say, they are the actual costs adjusted for differences in farm wages from a fixed base. In this form they serve as a basis for comparing the quantity of labor employed in one year with that employed in another. Thus the indications are that the quantity of labor employed per acre in 1928 was approximately 19 per cent. greater than the quantity employed in 1912.

Another illustration of when deflation may be desirable is in the instance of wages, since the wage rate itself does not clearly indicate its value in units of purchasing power. In 1913, for example, the daily farm wage without board in the United States was \$1.48,<sup>1</sup> whereas in 1928 it was \$2.43.<sup>1</sup> Considered alone, these wage rates might be misleading because of the cost of living factor. Even with the higher wage in 1928 the farm laborer may have been in no better financial circumstances than when cost of living was lower. In making an analysis of money wages, therefore, it may be advisable to compare purchasing power instead of the actual money wage.

<sup>1</sup> *Yearbook, United States Department of Agriculture, 1928*, page 1057, Table 531.

## CHAPTER XI

### LINEAR AND NONLINEAR RELATIONSHIPS

The discussion in Chapter X has made it easy to understand the meanings of linear and nonlinear relationships. These terms are so self-explanatory that only a brief mention will be made of them.

#### LINEAR RELATIONSHIP

Linear relationship is merely that relationship between corresponding pairs of items in a series that can be represented by a continuous straight line. For example, if we were plotting the ages of pupils and the classes to which they belong in school it would ordinarily be expected that for each year there is an increase in age there would be a corresponding promotion to the next higher grade. Thus, the relationship between age and class is not only linear, but positive linear, since a straight line with an upward slope shows that an increase of one unit in one series is accompanied by an increase of one unit in another. A decrease of one unit in a series accompanying a corresponding increase would be negative linear relationship. It will be understood, of course, that an increase in one series may be accompanied by a less than proportionate decrease in another and yet the relationship may be linear. The same may be said for a less than proportionate increase in one series that accompanies an increase in another, or unequal decreases that are associated. Wind velocity is directly related to the drying off of the soil.

This does not indicate, however, that if the velocity is doubled the evaporation of moisture is likewise doubled. Other factors may enter in to prevent this.

In Table LXIX are recorded the statistics of production and price of cotton in the United States for a period of years. A comparison of these two series will show that with the exception of the years 1922 and 1923 an increase in production was always accompanied by a decrease in price, while a decrease in production was associated with an increase in price. The cotton crops in 1922 and 1923, when an increase in production was associated in each of these years with an increase in price, were below the normal in the United States, so that it would ordinarily be expected that there would be an increase in price in each of these years. In all other years for which data are presented in Table LXIX there were movements in opposite directions. This opposite tendency is indicative of negative relationship, and the nature of the trends are

TABLE LXIX. PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1916-1928 <sup>1</sup>

| YEAR | PRODUCTION        | PRICE PER POUND<br>RECEIVED BY PRO-<br>DUCERS, DECEMBER 1 |
|------|-------------------|-----------------------------------------------------------|
|      | <i>1000 Bales</i> | <i>Cents</i>                                              |
| 1916 | 11,450            | 19.6                                                      |
| 1917 | 11,302            | 27.7                                                      |
| 1918 | 12,041            | 26.6                                                      |
| 1919 | 11,421            | 35.6                                                      |
| 1920 | 13,440            | 13.9                                                      |
| 1921 | 7,954             | 16.2                                                      |
| 1922 | 9,755             | 23.8                                                      |
| 1923 | 10,140            | 31.0                                                      |
| 1924 | 13,628            | 22.6                                                      |
| 1925 | 16,104            | 18.2                                                      |
| 1926 | 17,977            | 10.9                                                      |
| 1927 | 12,955            | 19.6                                                      |
| 1928 | 14,373            | 18.0                                                      |

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.



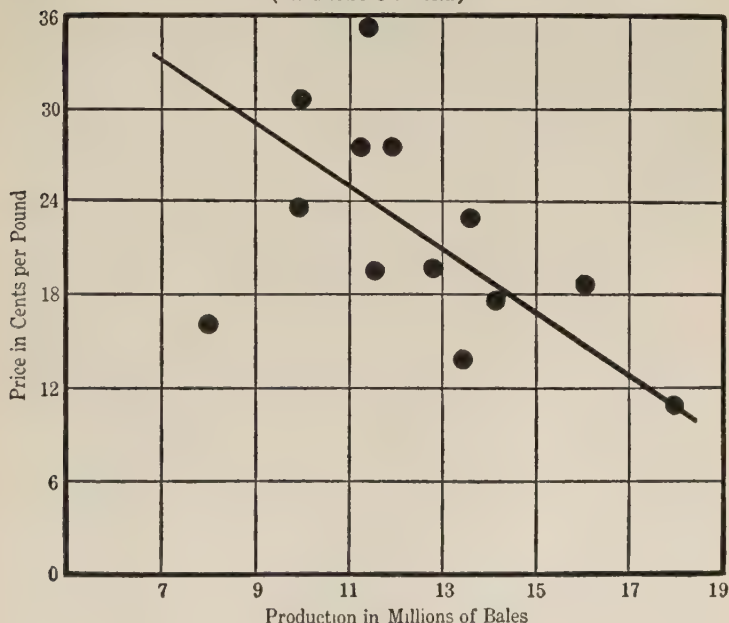
such that the relationship appears to be linear, or nearly linear.

The association between these two series, whose relationship is fairly linear, can best be shown by graphic presentation. This may be done by simply plotting the

CHART 36. LINEAR RELATIONSHIP BETWEEN PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1916-1928

(Based on data in Table LXIX)

(REGRESSION LINE)



data in an ordinary line chart, scatter diagram, or simple regression chart.<sup>1</sup> The latter is selected for this purpose. In Chart 36, therefore, we have a presentation of the straight line relationship between production and price as given in Table LXIX.

<sup>1</sup> See Chapter XIII for method of construction.

It will be observed that of the thirteen dots representing the points of intersection of pairs of items, six fall below the straight line and six fall above it. The thirteenth dot is divided by the line. It will be observed also that the line passes approximately through the center of the plotted area in the chart. This is linear relationship, the price of cotton decreasing with increased production and increasing with decreased production to such an extent that

TABLE LXX.    ACREAGE AND YIELD OF CORN ON SELECTED FARMS IN  
GEORGIA AND INDIANA IN 1928 <sup>1</sup>

| FARM NUMBER | ACREAGE PLANTED * | YIELD PER ACRE |
|-------------|-------------------|----------------|
|             | <i>Acres</i>      | <i>Bushels</i> |
| 1           | 11                | 34             |
| 2           | 14                | 28             |
| 3           | 18                | 47             |
| 4           | 2                 | 12             |
| 5           | 4                 | 9              |
| 6           | 18                | 39             |
| 7           | 3                 | 15             |
| 8           | 27                | 40             |
| 9           | 26                | 46             |
| 10          | 7                 | 30             |
| 11          | 23                | 49             |
| 12          | 28                | 38             |
| 13          | 8                 | 18             |
| 14          | 1                 | 9              |
| 15          | 5                 | 23             |
| 16          | 21                | 41             |
| 17          | 16                | 43             |
| 18          | 12                | 35             |
| 19          | 15                | 33             |
| 20          | 4                 | 13             |
| 21          | 26                | 35             |
| 22          | 28                | 45             |
| 23          | 13                | 40             |
| 24          | 24                | 43             |
| 25          | 29                | 38             |
| 26          | 12                | 25             |
| 27          | 1                 | 14             |

\* Acreage planted and acreage harvested were identical.

<sup>1</sup> Specially selected from data obtained by interviews in July, 1929.

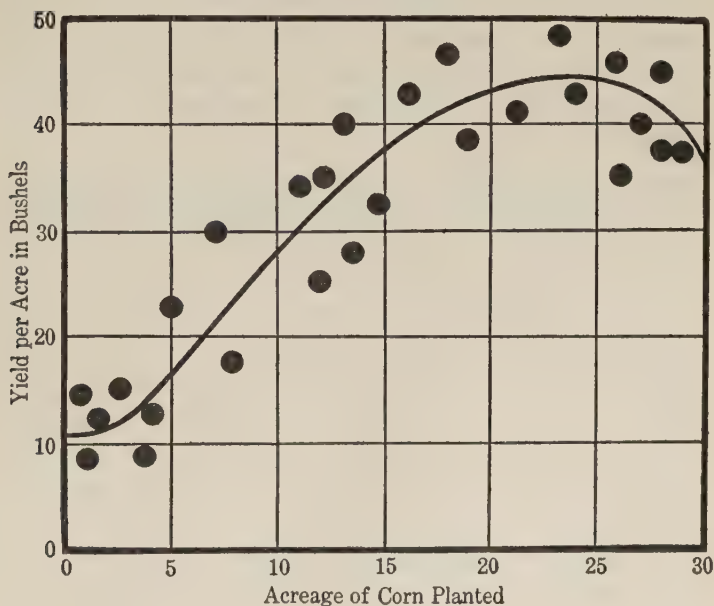
the movement of the two series in relation to each other can be represented by a continuous straight line.

### NONLINEAR RELATIONSHIP

Any relationship between corresponding pairs of items that cannot be represented by a straight line is said to be

CHART 37. NONLINEAR RELATIONSHIP BETWEEN YIELD OF CORN AND ACREAGE PLANTED, 1928

(Based on data in Table LXX)



nonlinear. Under actual conditions there are probably far more of these relationships than of the linear type.

In Table LXX are shown the statistics of acreage and yield of corn on twenty-seven farms in Georgia and Indiana in 1928. It was found that on the whole there was a

tendency for the higher yields to be associated with the larger acreages, but after a certain acreage was reached the yields tended to decline. The relationship between these two factors is such that it cannot properly be represented by a continuous straight line. It becomes necessary, therefore, to fit a curve of some other order to the data. Chart 37 shows this curve of relationship.

As will be very readily observed, a straight line would not properly show the relationship between yield and acreage. This is because yield does not increase proportionately with acreage in every instance.<sup>1</sup> The importance of ascertaining whether relationships are linear or nonlinear will be more fully realized when the subject of correlation is reviewed in the next chapter.

<sup>1</sup> The inference is not to be drawn that this relationship is necessarily true for the entire universe of which the sample is a part. The data represented in Chart 37 are specially selected, and are presented for illustrative purposes only.

## CHAPTER XII

### CORRELATION

In many statistical studies it is significant to determine whether or not there is any apparent cause and effect relationship between different series of items, and, if so, the extent of this relationship. Observations over a long period of time, or in a large number of instances, may show, for example, that there is a pronounced tendency for the volume of sales in retail stores to fluctuate in sympathy, or in the same direction, with unemployment, decreasing as employment decreases and increasing as employment increases. These changing conditions would indicate that the unemployment situation is exerting a causal influence upon the fluctuating volume of sales made by retail stores. Such a relationship is known as correlation, which may be defined as the mutual or reciprocal relationship existing between corresponding or associated items in two or more series. As is implied by the definition, in order for there to be correlation there must be evidences that the changes under consideration are the effects of a certain causal factor or group of factors. Increases and decreases in the importation of bananas, for example, may be associated with increases and decreases in the exportation of domestic wheat, but it would probably be difficult to show that either of these changing conditions is the cause or effect of the other. Such a relationship as this may be only coincidental, even though it persistently exists, and not correlation in the true sense.

As will be readily observed, whenever there is correlation there are two sets of conditions, those that cause or influence certain changes, and those that are the results or effects of those changes. In actual practice it is common to designate the causal factors as the independent series or variables, and the factors which are the results or effects as the dependent series or variables. The reason for these designations is obvious because, as suggested, one series, the series of causal factors, is independent of those changes effected by it. Thus the series with changes that are caused by this independent set of factors is the dependent series. Common usage has termed the independent items as the "x values" or variables and the dependent items as the "y values" or variables. Accordingly, in a case where changes in price are caused by changes in production the dependent series, price, becomes known as "y," and the independent series, production, is referred to as "x."

In introducing this general subject it is important to state clearly that correlation is of two kinds, positive and negative. Whenever changes in the independent series are accompanied or followed by corresponding changes in the same direction in the dependent series the indications are that the two series are positively related. On the other hand, if changes in the dependent series are accompanied or followed by opposite changes in the dependent series the conditions would seem to indicate a negative relationship. Changes in the independent that are equally likely to be accompanied or followed by changes in the same or opposite direction in the dependent series indicate an absence of correlation.

The fundamental principle, then, underlying correlation is that when two or more series are related there is a



tendency for a change in one to be invariably associated with either a like or opposite change in the other. To express this degree of relationship the coefficient of correlation has been devised, which is merely a mathematical measure showing how nearly or to what extent a change in the independent series is accompanied or followed by a change in the same or opposite direction of like magnitude in the dependent series. This measure, the correlation coefficient, which is represented by  $r$  is calculated by such method that it has a value of  $+1$  for perfect positive correlation and a value of  $-1$  for perfect inverse, or negative, correlation. Under actual conditions, of course, we seldom find a cause and effect relationship that measures as great as 1, so that the values of the coefficient of correlation are generally expected to be some fractional part of 1. Whether this fractional part will be a minus or plus quantity will depend upon the direction of correlation.

A question often arises as to how big a coefficient of correlation must be before it has significance. While there is much difference of opinion regarding this particular phase of the subject, depending somewhat, perhaps, upon the nature of the data correlated, there are certain definite rules that have been suggested for the interpretation of the coefficient. These suggestions are as follows: <sup>1</sup>

1. "If  $r$  is less than the probable error there is no evidence whatever of correlation."
2. "If  $r$  is more than six times the size of the probable error the existence of correlation is a practical certainty."
3. When the probable error is relatively small, "if  $r$  is less than .30 the correlation cannot be considered at all marked."

<sup>1</sup> W. I. King, *Elements of Statistical Method*, 1923, page 215.

4. If the probable error is relatively small, a coefficient "above .50 indicates decided correlation."

To show more clearly how a change in the items in one series may be expected to cause a change in corresponding items in another series, let us examine the statistics of cotton production and cotton prices in the following table. It will be seen that when the production decreased in 1917, for instance, there was an increase in the price for that year from 19.6 cents to 27.7 cents per pound, whereas an increase in production in 1924 over 1923 was associated with a price in 1924 which was less than that of the immediately preceding year. A check of corresponding changes for the other years will show that whenever production increased there was almost always a decrease in price, and that when there was a decrease in production the prices moved in the opposite direction.

TABLE LXXI. PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1916-1928 <sup>1</sup>

| YEAR | PRODUCTION         | PRICE PER POUND<br>RECEIVED BY PRO-<br>DUCERS, DECEMBER 1 |
|------|--------------------|-----------------------------------------------------------|
|      | <i>1,000 Bales</i> | <i>Cents</i>                                              |
| 1916 | 11,450             | 19.6                                                      |
| 1917 | 11,302             | 27.7                                                      |
| 1918 | 12,041             | 27.6                                                      |
| 1919 | 11,421             | 35.6                                                      |
| 1920 | 13,440             | 13.9                                                      |
| 1921 | 7,954              | 16.2                                                      |
| 1922 | 9,755              | 23.8                                                      |
| 1923 | 10,140             | 31.0                                                      |
| 1924 | 13,628             | 22.6                                                      |
| 1925 | 16,104             | 18.2                                                      |
| 1926 | 17,977             | 10.9                                                      |
| 1927 | 12,955             | 19.6                                                      |
| 1928 | 14,373             | 18.0                                                      |

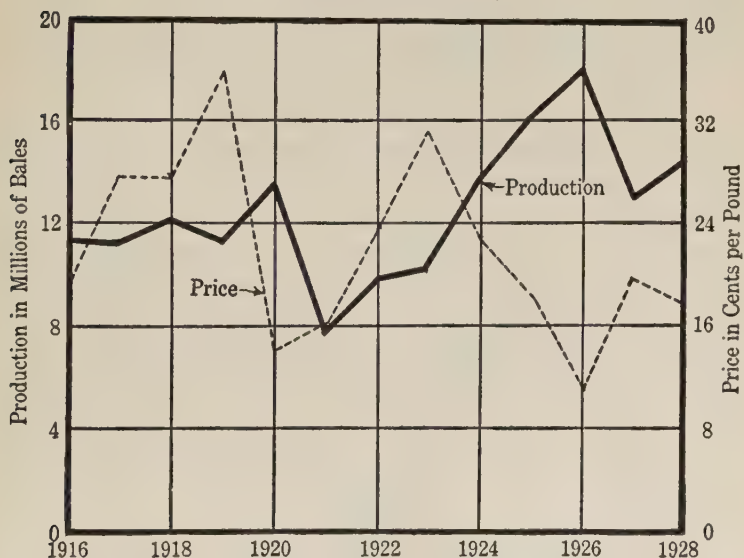
<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.

A better way to show this relationship between independent and dependent variables is to present the two series graphically, so that a direct comparison can be made of the movements or changes in one series in relation to the

CHART 38. PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1916-1928

(Based on data in Table LXXI)

(INVERSE RELATIONSHIP)

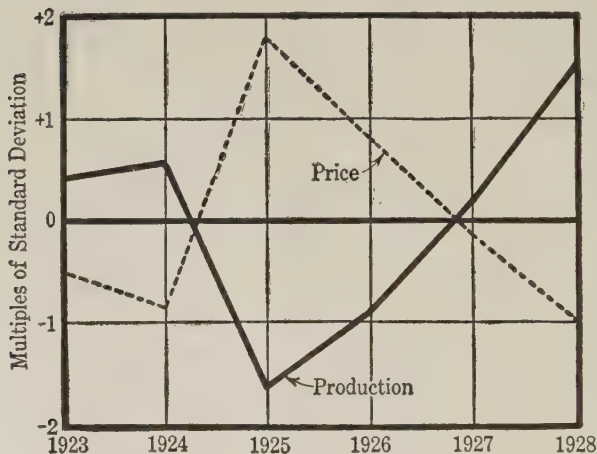


movements or changes in the other. Accordingly, in Chart 38 the production and price of cotton are plotted for the thirteen-year period, the continuous line representing production and the dotted line representing price. It can now be more readily seen that there is a pronounced tendency for the curves of the two series to move in opposite directions. That is to say, when the production of cotton increases there is usually a decrease in subsequent price, and

when there is a decrease in production there is an increase in price. This tendency for changes to fluctuate in opposite directions is indicative of inverse, or negative, correlation. If increases and decreases in production were associated consistently with corresponding increases and decreases in price in the same direction, the indications would be that there is positive correlation between the series.

CHART 39. INVERSE RELATIONSHIP BETWEEN PRODUCTION AND PRICE OF POTATOES IN THE UNITED STATES, 1923-1928

(Based on data in Table LXXIV)



For some purposes it may be satisfactory to make a graphic comparison of deviations from the arithmetic mean. A better comparison than this, however, would be to convert the deviations from the mean to multiples of the standard deviation and then plot these multiples on cross-section paper. The plotting of multiples of standard deviation is illustrated in Chart 39.

As will be observed in Table LXXIV, on which the chart is based, the multiples of standard deviation are calculated from the deviations of actual values from their mean. The multiples may be calculated also from first differences and from percentage changes of first differences, and they may even be derived from the actual values themselves.

Sometimes correlation is spoken of as the "typical" amount of negative or positive similarity existing between corresponding items in different series. In this connection it is important to note that the term "typical" has significance because the numerical expressions of relationship as calculated may not represent actualities. As an illustration of this, let us refer to a study made of the relationships between weather factors and yield of cotton per acre in Wake County, North Carolina, for the period from 1900 to 1927, inclusive. In 1915 the precipitation for the four-day period from July 16 to July 19 at a selected station represented an increase of 3,800 per cent. over the same period of the preceding year, and for the seven-day period from August 20 to August 26 in 1916 the increase in precipitation at the same station was 13,350 per cent. greater than that of the corresponding seven-day period in 1915. In such instances as these it is obvious that no method of correlation would show normal relationship unless adjustments were made for variations or unless additional factors be taken into consideration. It is, therefore, the duty of the investigator to study his data for probable cause and effect and to comprehend the significance of abnormally high and low magnitudes in any of the items in either series of variables.

In the last part of this chapter the construction of the correlation table is outlined to show how the relationships

between corresponding pairs of items in a frequency distribution are indicated.

### SIMPLE CORRELATION

When only one series is correlated with another, so that we measure only the relationship between one independent variable and a dependent variable, the correlation is said to be of the simple type. Thus the calculation of relationship between production and price of beef cattle would be simple correlation because only production of beef cattle, the independent variable, is considered in relation to price, the dependent variable. The more commonly used methods of simple correlation will be discussed.

TABLE LXXII. PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1917-1928 <sup>1</sup>

| 1    | 2                                                        | 3                                                        |
|------|----------------------------------------------------------|----------------------------------------------------------|
| YEAR | PRODUCTION                                               | PRICE PER POUND<br>RECEIVED BY<br>PRODUCERS, DECEMBER 1  |
|      | First Differences<br>(Deviations from<br>Preceding Year) | First Differences<br>(Deviations from<br>Preceding Year) |
|      | <i>1,000 Bales</i>                                       | <i>Cents</i>                                             |
| 1917 | — 148                                                    | 8.1                                                      |
| 1918 | 739                                                      | — .1                                                     |
| 1919 | — 620                                                    | 8.0                                                      |
| 1920 | 2,019                                                    | — 21.7                                                   |
| 1921 | — 5,486                                                  | 2.3                                                      |
| 1922 | 1,801                                                    | 7.6                                                      |
| 1923 | 385                                                      | 7.2                                                      |
| 1924 | 3,488                                                    | — 8.4                                                    |
| 1925 | 2,576                                                    | — 4.4                                                    |
| 1926 | 1,873                                                    | — 7.3                                                    |
| 1927 | — 5,022                                                  | 8.7                                                      |
| 1928 | 1,418                                                    | — 1.6                                                    |

<sup>1</sup> Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.



*Coefficient of Concurrent or Concomitant Deviations*

If it is desired to know only whether two series move in the same direction, or if one series moves in the opposite direction from the other, the concurrent or concomitant deviations may be used as a basis for measurement. Concurrent or concomitant deviations are those deviations that are in the same directions for corresponding items in each series. Thus, in Table LXXII it will be seen that in 1922 and 1923 the deviations of both production and price are plus quantities. These, then, are concurrent deviations.

The following formula has been found very convenient for calculating the coefficient of concurrent deviations:

$$r = \pm \sqrt{\frac{2c - n}{n}} \quad ^1$$

In this formula the meanings of the symbolic expressions are as follows:

$r$  = the coefficient of correlation

$c$  = the number of concurrent deviations

$n$  = the number of pairs of items.

Referring again to Table LXXII it will be seen that there are twelve pairs of items. In all instances but two, a plus or minus deviation in one series is associated with an opposite deviation in the other. For the years 1922 and 1923, as has been stated, the deviations in both series are plus quantities, and since they are thus alike they

<sup>1</sup> W. I. King, *Elements of Statistical Method*, 1923, pages 207-211.

are concurrent. Applying the formula given above we have:

$$r = \frac{+ \sqrt{\frac{2(2) - 12}{n}}}{n}$$

$$= \sqrt{\frac{8}{12}} = -.816$$

It will be noted that when the number of pairs of items, 12, is subtracted from two times the number of concurrent deviations we have a value of  $-8$ . Before the square root is extracted this minus sign must be removed. The sign is again added, however, when the square root is extracted.<sup>1</sup>

The principal value of the coefficient of concurrent deviations is that it indicates the direction of the movement of one series in relation to another. As has been seen, it is based entirely upon the direction of deviations of the variables and is in no way related to the magnitude of these deviations. It may be said, therefore, that by the method of concurrent deviations all the deviations are assumed to be identical in size. For this reason a perfect correlation would be obtained whenever the variables in the two series move consistently in the same or opposite directions. For instance, in Table LXXII if all the associated deviations in columns 2 and 3 were either of minus or plus value we would have twelve concurrent deviations, in which case there would be perfect positive correlation. On the other hand, if deviations in column 2

<sup>1</sup> The square root of any number may be extracted by the use of logarithms, which is often convenient if a table of logarithms is available. The method consists merely of looking up the logarithm of the number, dividing the logarithm by 2, and then finding the antilogarithm of the quotient. For example, in extracting the square root of 144 we look up the logarithm of 144, which is 2.1584. Dividing 2.1584 by 2 gives 1.0792. Looking up the antilogarithm of .0792 we find it to be 12, the square root of 144.

were always associated with deviations in the opposite direction in column 3 there would be perfect negative correlation. It will be seen, therefore, that great limitation must be placed upon the significance of the coefficient of concurrent deviations.

This discussion naturally leads to the necessity of considering some methods of correlation by which the size as well as the direction of the deviations is measured. The first of these measures to be considered is the Pearsonian coefficient, which is based on the extent to which corresponding pairs of items deviate from their respective simple arithmetic means.

### *Pearsonian Coefficient*

In much of the ordinary statistical analysis the Pearsonian method of correlation is used. This method has become known also as the product-moment method and as the sum-product method. Being originally devised to remedy the weakness of the method of concurrent deviations, it has since become very popular, largely, perhaps, because the calculations involved are very simple and easily understood. It is chiefly on account of the latter factor that the Pearsonian method of correlation is one of the most commonly used in the analysis of economic data. For the calculation of the coefficient the following formula has been found to be very convenient:

$$r = \frac{\Sigma xy}{n(S.D.x) (S.D.y)},$$

which becomes 
$$r = \frac{\Sigma xy}{n \sigma x \sigma y}.$$

The meanings of the symbolic expressions are as follows:

$r$  = the coefficient of correlation

$\Sigma xy$  = the summation of the products of corresponding pairs of deviations from the means.<sup>1</sup>

TABLE LXXIII. PRODUCTION AND PRICE OF POTATOES IN THE UNITED STATES, 1923-1928 <sup>2</sup>

(Calculation of the Pearsonian Coefficient of Correlation)

| 1     | 2                 | 3                                 | 4                              | 5                     | 6                                   | 7                              | 8                                                         |
|-------|-------------------|-----------------------------------|--------------------------------|-----------------------|-------------------------------------|--------------------------------|-----------------------------------------------------------|
| YEAR  | PRODUCTION        |                                   |                                | PRICE                 |                                     |                                | PRODUCT OF DEVIATIONS FROM MEAN (Column 3 Times Column 6) |
|       | TOTAL             | DEVIATIONS FROM MEAN (Mean = 397) | SQUARE OF DEVIATIONS FROM MEAN | PER BUSHEL DECEMBER 1 | DEVIATIONS FROM MEAN (Mean = 103.2) | SQUARE OF DEVIATIONS FROM MEAN |                                                           |
|       |                   | $x$                               | $x^2$                          |                       | $y$                                 | $y^2$                          |                                                           |
|       | 1,000,000 Bushels | 1,000,000 Bushels                 |                                | Cents                 | Cents                               |                                |                                                           |
| 1923  | 416               | 19                                | 361                            | 78.1                  | -25.1                               | 630.01                         | - 476.9                                                   |
| 1924  | 422               | 25                                | 625                            | 62.5                  | -40.7                               | 1656.49                        | - 1017.5                                                  |
| 1925  | 323               | -74                               | 5,476                          | 186.8                 | 83.6                                | 6988.96                        | - 6186.4                                                  |
| 1926  | 354               | -43                               | 1,849                          | 141.4                 | 38.2                                | 1459.24                        | - 1642.6                                                  |
| 1927  | 403               | 6                                 | 36                             | 96.5                  | - 6.7                               | 44.89                          | - 40.2                                                    |
| 1928  | 463               | 66                                | 4,356                          | 54.0                  | -49.2                               | 2420.64                        | - 3247.2                                                  |
| Total | —                 | *                                 | 12,703                         | —                     | *                                   | 13,200.23                      | -12,610.8                                                 |

\* If fractions were carried, the total of this column would be zero.

<sup>1</sup> The use of logarithms in multiplication is sometimes found to be very convenient. To find the product of two numbers we merely look up their logarithms, summate these logarithms, look up the number corresponding to the mantissa, or fractional part of the sum of the two logarithms, and then move the decimal point to the right as many places as it takes to give the product one more digit than the size of the characteristic of the sum of the two logarithms. For example, let us illustrate how the product of 14 times 16 is obtained by the use of logarithms. The logarithm of 14 is 1.1461 and of 16 it is 1.2041. The summation of these two logarithms is 2.3502. The number corresponding to .3502, the fractional part of the summation, is 224. We know the product of 14 and 16 must contain one digit more than the size of the characteristic of the summation of the two logarithms. This summation, as we have seen, is 2.3502. Since the characteristic is 2, the number of digits in the product of 14 and 16 must have one digit more than 2, or 3 digits. Therefore, the product is 224.

<sup>2</sup> Yearbook, United States Department of Agriculture, 1928, page 806, Table 202.

$n$  = the number of pairs of items

$\sigma x$  = the standard deviation of the  $x$  series

$\sigma y$  = the standard deviation of the  $y$  series.

By means of this calculation, weight is given to extreme deviations when the standard deviations are determined. This is partly the reason the value of  $n\sigma x\sigma y$  is greater than the value  $\Sigma xy$ .

The preceding table will be used in illustrating the calculation of the coefficient of correlation from the above formula.

Proceeding with the calculation and substituting the proper values we have:

$$r = \frac{\Sigma xy}{n\sigma x\sigma y} = \frac{-12,610.8}{6 \times 46.0 \times 46.9} = -.974$$

Perhaps it will be well to inquire into the exact meaning of this coefficient of correlation. The significance of the coefficient lies in the fact that it is a measurement of the size or magnitude of pairs of deviations from the arithmetic mean. Therefore, the value of  $-.974$  is an indication of the extent to which fluctuations in production of potatoes are accompanied by fluctuations in price. Squaring  $-.974$  gives us 95. This is considered to indicate that about 95 per cent. of the fluctuations in production are accompanied or followed by corresponding fluctuations in price in the opposite direction,<sup>1</sup> and that 5 per cent. of price fluctuations are independent of fluctuations in production, or at least do not accompany fluctuations in production. Hence it may be possible to obtain a correlation as high as .22 between price and factors not associ-

<sup>1</sup> This relationship may be considered also to indicate that about 95 per cent. of the changes in price are caused by changes in production.

ated with production. This value of .22 is obtained by subtracting .95 from 1.00 and then extracting the square root of the remainder.<sup>1</sup>

Although the Pearsonian method of correlation is quite satisfactory in many instances, it has certain important weaknesses or limitations that must be considered. Let us assume that we have two long-time historical series that show a decided trend in the same direction for the period as a whole, and in which the year to year changes are in opposite directions. Obviously, the Pearsonian method of correlation would show a positive long-time relationship, but would not take into account the negative relationships between corresponding pairs of items for the shorter periods.

It is for this reason that under any such similar circumstance the Pearsonian coefficient may prove unsatisfactory, particularly if interest lies in short-time changes. Another serious weakness, or rather another limitation, of the Pearsonian coefficient is that its calculation is based on the assumption that the relationship between variables in the two series is linear, while, as a matter of fact, such relationships may not often occur, especially with economic data.

### *Coefficient of Correlation from Multiples of Standard Deviation*

By the Pearsonian method of correlation, the summation of the products of corresponding pairs of deviations from the mean is divided by the product of the number of pairs of deviations and the two standard deviations to obtain the coefficient of correlation. Another very satisfactory

<sup>1</sup> The explanation in this paragraph applies to all coefficients of correlation, and so will not be repeated as the several methods of correlation are discussed.



way by which the coefficient of correlation may be calculated consists of dividing the deviations by the standard deviation, which gives the multiples of standard deviation, obtaining the products of corresponding pairs of multiples, summing the products algebraically, and then dividing by the number of pairs of multiples. Let us illustrate the method by using the same data as presented in Table LXXIII.

TABLE LXXIV. PRODUCTION AND PRICE OF POTATOES IN THE UNITED STATES, 1923-1928 <sup>1</sup>

(Calculation of the Coefficient of Correlation from Multiples of Standard Deviation of Deviations from the Mean of Actual Values)

| 1     | 2                                              | 3                                                                             | 4                                             | 5                                                                             | 6                                                                                      |
|-------|------------------------------------------------|-------------------------------------------------------------------------------|-----------------------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------------|
| YEAR  | PRODUCTION                                     |                                                                               | PRICE                                         |                                                                               | PRODUCT OF<br>MULTIPLES<br>OF STANDARD<br>DEVIATION<br>(Column 3<br>Times<br>Column 5) |
|       | DEVIATIONS<br>FROM THE<br>MEAN<br>(Mean = 397) | MULTIPLES OF<br>THE STANDARD<br>DEVIATION<br>(Standard De-<br>viation = 46.0) | DEVIATIONS<br>FROM THE MEAN<br>(Mean = 103.2) | MULTIPLES OF<br>THE STANDARD<br>DEVIATION<br>(Standard De-<br>viation = 46.9) |                                                                                        |
|       | dm                                             | x                                                                             | dm                                            | y                                                                             | xy                                                                                     |
|       | 1,000,000<br>Bushels                           |                                                                               | Cents                                         |                                                                               |                                                                                        |
| 1923  | 19                                             | .4130                                                                         | -25.1                                         | -.5352                                                                        | -.2210                                                                                 |
| 1924  | 25                                             | .5435                                                                         | -40.7                                         | -.8678                                                                        | -.4716                                                                                 |
| 1925  | -74                                            | -1.6087                                                                       | 83.6                                          | 1.7825                                                                        | -2.8675                                                                                |
| 1926  | -43                                            | -.9348                                                                        | 38.2                                          | .8145                                                                         | -.7614                                                                                 |
| 1927  | 6                                              | .1304                                                                         | -6.7                                          | -.1429                                                                        | -.0186                                                                                 |
| 1928  | 66                                             | 1.4348                                                                        | -49.2                                         | -1.0490                                                                       | -1.5051                                                                                |
| Total | —                                              | —                                                                             | —                                             | —                                                                             | -5.8452                                                                                |

The multiples of standard deviation in the preceding table, columns 3 and 5, are obtained by dividing the standard deviation of each series into the deviations from the mean. Values in column 3, for example, are obtained by dividing 46.0, the standard deviation of the x series as calculated from column 4 of Table LXXIII, into the devia-

<sup>1</sup> Data from Table LXXIII, page 206.

tions of the  $x$  items from their mean. The multiples of standard deviation of price in column 5 are obtained by dividing 46.9, the standard deviation of the  $y$  series, into the deviations of the  $y$  items from their mean.<sup>1</sup> The multiples of standard deviation, then, in columns 3 and 5, are simply the deviations from the mean expressed in terms of the standard deviation. It is obvious, therefore, that by multiplying these multiples by the standard deviation the original deviations are obtained.

Column 6 shows the products of corresponding pairs of multiples, with a summation of  $-5.8452$ . Applying the formula for the coefficient of correlation we have:

$$r = \frac{\Sigma xy}{n} = \frac{-5.8452}{6} = -.974,$$

the same as the coefficient by the Pearsonian method.

The essential feature to remember in regard to  $\Sigma xy$ , summation of column 6, is that the products are to be added algebraically. The products carry the signs of the multiples of standard deviation. It so happens that the signs of the pairs of multiples in Table LXXIV are unlike, so that all the products in column 6 carry the minus sign. If some of the products were plus quantities they would have been summated in one column, the minus quantities in another, and the difference between the two summations obtained. This difference, either a plus or a minus quantity, would have been the total summation of  $xy$ , or column 6, and it would have been divided

<sup>1</sup> Instead of dividing each individual item by the standard deviation it is a good plan, particularly if calculating machines are available, to multiply the reciprocal of the standard deviation by the number of which the multiple is being calculated. The reciprocal of the standard deviation, or of any other number or value as to that matter, is obtained by dividing the number into 1. Thus, in Table LXXIV if we divide 46.0, the standard deviation of production, the  $x$  series, into 1 and then multiply by the deviations in column 2 the multiples of standard deviation in column 3 are obtained.

by the total number of pairs of multiples of standard deviation to obtain the coefficient of correlation.

### *Coefficient of Correlation from Actual Values*

If the individual items are not too large the coefficient of correlation may be conveniently calculated without determining the deviation of items from the mean. This method consists of using the actual  $x$  and  $y$  values and then employing the means as correction factors.

The first step is to calculate the means of each series. This, of course, is done by merely summing each series and dividing by the number of items. The standard deviation is then determined by squaring the actual values, dividing by the number of items, subtracting the square of the mean, and then extracting the square root. The formula for the standard deviation by this method is

$$\sigma = \sqrt{\frac{\sum d^2}{n} - c^2}.$$

The mean of the series is the correction factor because in squaring the actual values we assume the mean to be zero and the actual values or items in the series to be the deviations.

Next the summation of the products of corresponding items in the two series is obtained. This is done by multiplying each item of the  $x$  series by the associated item of the  $y$  series and then adding the individual products to obtain the total summation. From this summation the product-moment is calculated by the formula

$$p(\text{product-moment}) = \frac{\sum xy}{n} - M_x M_y,$$

in which  $\sum xy$  is the summation of the products of corresponding pairs of items,  $n$  the number of pairs of items,  $M_x$

the mean of the x series, and  $M_y$  the mean of the y series. The product-moment thus obtained is divided by the product of the two standard deviations to obtain the coefficient of correlation. The formula for the coefficient of correlation, therefore, is

$$r = \frac{\frac{\Sigma xy}{n} - M_x M_y}{\sigma_x \sigma_y}.$$

With fractions carried to a reasonable place this method of correlation will give approximately the same result as the Pearsonian method and the method as illustrated in Table LXXIV.

### *Correlation Based on Deviations from Straight Line Trend*

If we are interested in measuring the correlation between two series that are represented by a straight line trend of least squares we do not base our calculation on deviations from the mean, but on deviations from trend. By this method the standard deviation of each series is based on the summation of the squares of deviations from trend, while the xy summation, corresponding to the summation of column 8 in Table LXXIII, is the total of the products obtained by multiplying together the deviations of items in one series from the trend and the deviations in the other.

Calculation of the coefficient of correlation by this method may be based on the deviations of actual values in each series from the trend, or the secular trend and seasonal variations, if any, may be removed before the two series are compared. For most purposes in the analyzing of economic data it is probably better to base the co-

efficient of correlation on deviations from trend in actual values without making adjustments for secular trend and seasonal variations.

This method of calculating the coefficient of correlation, as implied, can be used only when there is linear relationship between the two series. When the relationship is nonlinear, another measure, the correlation index, is used to show how well a certain curve indicates the relationship between pairs of variables in two series, or how well the particular curve represents the data.<sup>1</sup> The correlation index is discussed later in this chapter.

### *Correlation of First Differences*

Sometimes in comparing fluctuations or changes in one series with fluctuations or changes in another we are interested in knowing the extent to which a change in a variable over the immediately preceding one is associated with a like or opposite change in a variable of another series. For example, if the production of cotton in 1927 in the United States were five million bales less than in 1926 it may be desired to know the extent of change in the price of 1927 as compared with the price in 1926. Accordingly, we would not measure this relationship on the basis of deviations from means, as was done by the Pearsonian method, but from deviations of individual items from immediately preceding items. These deviations are known as first differences, from which this method of correlation derives its name. The calculation of this coefficient of correlation is illustrated in the following table.

<sup>1</sup> The coefficient of correlation calculated from deviations from straight line trend has sometimes been termed an index of correlation. It is probably better, however, to refer to it as a coefficient of correlation because a straight line trend of least squares fitted to a series passes through the mean, and the regression line showing highest relationship between variables in two series passes through the means of both series.



TABLE LXXV. PRODUCTION AND PRICE OF  
(Calculation of the Coefficient of Cor-

| 1     | 2                                                          | 3                                             | 4                                        |
|-------|------------------------------------------------------------|-----------------------------------------------|------------------------------------------|
| YEAR  | PRODUCTION                                                 |                                               |                                          |
|       | FIRST DIFFERENCES,<br>OR DEVIATIONS FROM<br>PRECEDING YEAR | DEVIATIONS FROM<br>THE MEAN<br>(Mean = 2,436) | SQUARE OF<br>DEVIATIONS FROM<br>THE MEAN |
|       | x                                                          | dmx                                           | dmx <sup>2</sup>                         |
|       | <i>1,000 Bales</i>                                         |                                               |                                          |
| 1917  | — 148                                                      | —2,584                                        | 6,677,056                                |
| 1918  | 739                                                        | —1,697                                        | 2,879,809                                |
| 1919  | — 620                                                      | —3,056                                        | 9,339,136                                |
| 1920  | 2,019                                                      | — 417                                         | 173,889                                  |
| 1921  | —5,486                                                     | —7,922                                        | 62,758,084                               |
| 1922  | 1,801                                                      | — 635                                         | 403,225                                  |
| 1923  | 385                                                        | —2,051                                        | 4,206,601                                |
| 1924  | 3,488                                                      | 1,052                                         | 1,106,704                                |
| 1925  | 2,476                                                      | 40                                            | 1,600                                    |
| 1926  | 1,873                                                      | — 563                                         | 316,969                                  |
| 1927  | —5,022                                                     | —7,458                                        | 55,621,764                               |
| 1928  | 1,418                                                      | —1,018                                        | 1,036,324                                |
| Total | —                                                          | —                                             | 144,521,161                              |

$$\sigma x = \frac{\sqrt{\Sigma d^2}}{n} = 3,470$$

By this method the standard deviation is calculated from the sum of the squares of deviations from the mean of first differences. These deviations are shown in columns 3 and 6. In column 8 are the products of these deviations, the summation of which is divided by the product of the number of pairs of deviations, or items, and the two standard deviations to obtain the coefficient of correlation. The summation of column 8 is —190,644.9, and the product of the number of pairs of deviations, or items, and the two standard deviations (12 times 3,470 times 8.96) is 373,094.4. The quotient obtained by division is —.511, the coefficient of correlation of first differences.



COTTON IN THE UNITED STATES, 1917-1928 <sup>1</sup>  
relation from First Differences)

| 5                                                          | 6                                            | 7                                        | 8                                                                         |
|------------------------------------------------------------|----------------------------------------------|------------------------------------------|---------------------------------------------------------------------------|
| PRICE                                                      |                                              |                                          | PRODUCT OF<br>DEVIATIONS<br>FROM THE MEAN<br>(Column 3 Times<br>Column 6) |
| FIRST DIFFERENCES, OR<br>DEVIATIONS FROM<br>PRECEDING YEAR | DEVIATIONS<br>FROM THE MEAN<br>(Mean = -1.6) | SQUARE OF<br>DEVIATIONS FROM<br>THE MEAN |                                                                           |
| y                                                          | dmy                                          | dmy <sup>2</sup>                         | dmx.dmy (xy)                                                              |
| <i>Cents per Pound</i>                                     |                                              |                                          |                                                                           |
| 8.1                                                        | 9.7                                          | 94.09                                    | -25,064.8                                                                 |
| - .1                                                       | 1.5                                          | 2.25                                     | - 2,545.5                                                                 |
| 8.0                                                        | 9.6                                          | 92.16                                    | -29,337.6                                                                 |
| -21.7                                                      | -20.1                                        | 404.01                                   | 8,381.7                                                                   |
| 2.3                                                        | 3.9                                          | 15.21                                    | -30,895.8                                                                 |
| 7.6                                                        | 9.2                                          | 84.64                                    | - 5,842.0                                                                 |
| 7.2                                                        | 8.8                                          | 77.44                                    | -18,048.8                                                                 |
| - 8.4                                                      | - 6.8                                        | 46.24                                    | - 7,153.6                                                                 |
| - 4.4                                                      | - 2.8                                        | 7.84                                     | - 112.0                                                                   |
| - 7.3                                                      | - 5.7                                        | 32.49                                    | - 3,209.1                                                                 |
| 8.7                                                        | 10.3                                         | 106.09                                   | -76,817.4                                                                 |
| - 1.6                                                      | 0                                            | 0                                        | 0                                                                         |
| —                                                          | —                                            | —                                        | -190,644.9                                                                |

$$\sigma y = \sqrt{\frac{\sum d^2}{n}} = 8.96$$

$$r = \frac{\sum dmx.dmy \text{ (or } \sum xy)}{n \sigma x \sigma y} = -0.511$$

The first difference method is of particular value in making the analysis of relationship between corresponding pairs of items in two long-time historical series. As has already been indicated in the discussion of the Pearsonian method of correlation, the long-time trend of such historical series may be in the same direction, while the short-time changes in the two series may be negatively related. Since the first difference method is based on successive corresponding changes in the items over the ones immediately preceding, the calculation of the co-

<sup>1</sup> Data taken from Table LXXII, page 202.

TABLE LXXVI. PRODUCTION AND PRICE OF  
(Correlation by the Method of Per-

| 1     | 2                                 | 3                                                                    | 4                                               | 5                                                                       | 6                                                                 | 7                                       |
|-------|-----------------------------------|----------------------------------------------------------------------|-------------------------------------------------|-------------------------------------------------------------------------|-------------------------------------------------------------------|-----------------------------------------|
| YEAR  | PRODUCTION                        |                                                                      |                                                 |                                                                         |                                                                   |                                         |
|       | Total in<br>Thousands<br>of Bales | First Differ-<br>ences, or<br>Deviations<br>from Pre-<br>ceding Year | Percentage<br>Change of<br>First<br>Differences | Deviation<br>of Percent-<br>age Change<br>from Mean<br>(Mean =<br>4.23) | Square of<br>Deviations<br>of Percent-<br>age Change<br>from Mean | Multiples<br>of Standard<br>Deviation * |
|       | $\Sigma$                          | $\Sigma x - x$                                                       | $\frac{\Sigma x - x}{x}$                        | $dmx$                                                                   | $dmx^2$                                                           | $MSDx$                                  |
| 1916  | 11,450                            |                                                                      |                                                 |                                                                         |                                                                   |                                         |
| 1917  | 11,302                            | — 148                                                                | — 1.29                                          | — 5.52                                                                  | 30.47                                                             | — .273                                  |
| 1918  | 12,041                            | 739                                                                  | 6.54                                            | 2.31                                                                    | 5.34                                                              | .114                                    |
| 1919  | 11,421                            | — 620                                                                | — 5.15                                          | — 9.38                                                                  | 87.98                                                             | — .464                                  |
| 1920  | 13,440                            | 2,019                                                                | 17.68                                           | 13.45                                                                   | 180.90                                                            | .665                                    |
| 1921  | 7,954                             | — 5,486                                                              | — 40.82                                         | — 45.05                                                                 | 2029.50                                                           | — 2.227                                 |
| 1922  | 9,755                             | 1,801                                                                | 22.64                                           | 18.41                                                                   | 338.93                                                            | .910                                    |
| 1923  | 10,140                            | 385                                                                  | 3.95                                            | — .28                                                                   | .08                                                               | — .014                                  |
| 1924  | 13,628                            | 3,488                                                                | 34.40                                           | 30.17                                                                   | 910.23                                                            | 1.491                                   |
| 1925  | 16,104                            | 2,476                                                                | 18.17                                           | 13.94                                                                   | 194.32                                                            | .689                                    |
| 1926  | 17,977                            | 1,873                                                                | 11.63                                           | 7.40                                                                    | 54.76                                                             | .366                                    |
| 1927  | 12,955                            | — 5,022                                                              | — 27.94                                         | — 32.17                                                                 | 1034.91                                                           | — 1.590                                 |
| 1928  | 14,373                            | 1,418                                                                | 10.95                                           | 6.72                                                                    | 45.16                                                             | .332                                    |
| Total | —                                 | —                                                                    | —                                               | —                                                                       | 4912.58                                                           | —                                       |

\* Standard deviation of  $x$  ( $S.D.x$ ) = 20.23

efficient of correlation by this method accounts for short-time changes.

*Correlation by the Method of Percentage Change of  
First Differences*

We have just seen how the coefficient of correlation is calculated from first differences. By a modification of that method the first differences are converted to terms of percentages, the standard deviation calculated from the summation of the squares of the deviations of percentage changes from their mean, and the deviations of percentage changes from their mean expressed as mul-

COTTON IN THE UNITED STATES, 1916-1928 <sup>1</sup>

centage Change of First Differences)

| 8                                                      | 9                                                    | 10                                     | 11                                                     | 12                                                  | 13                               | 14                                                                     |
|--------------------------------------------------------|------------------------------------------------------|----------------------------------------|--------------------------------------------------------|-----------------------------------------------------|----------------------------------|------------------------------------------------------------------------|
| PRICE                                                  |                                                      |                                        |                                                        |                                                     |                                  |                                                                        |
| Price in Cents per Pound Received by Producers, Dec. 1 | First Differences, or Deviations from Preceding Year | Percentage Change of First Differences | Deviation of Percentage Change from Mean (Mean = 7.31) | Square of Deviations of Percentage Change from Mean | Multiples of Standard Deviation* | Products of Multiples of Standard Deviation (Column 7 Times Column 13) |
| y                                                      | y - y                                                | $\frac{y - y}{y}$                      | dmy                                                    | dmy <sup>2</sup>                                    | MSDy                             | MSDx.<br>MSDy                                                          |
| 19.6                                                   |                                                      |                                        |                                                        |                                                     |                                  |                                                                        |
| 27.7                                                   | 8.1                                                  | 41.33                                  | 34.02                                                  | 1157.36                                             | .874                             | — .239                                                                 |
| 27.6                                                   | — .1                                                 | — .36                                  | — 7.67                                                 | 58.83                                               | — .197                           | — .022                                                                 |
| 35.6                                                   | 8.0                                                  | 28.99                                  | 21.68                                                  | 470.02                                              | .557                             | — .258                                                                 |
| 13.9                                                   | — 21.7                                               | — 60.96                                | — 68.27                                                | 4660.79                                             | — 1.754                          | — 1.166                                                                |
| 16.2                                                   | 2.3                                                  | 16.55                                  | 9.24                                                   | 85.38                                               | .237                             | — .528                                                                 |
| 23.8                                                   | 7.6                                                  | 46.91                                  | 39.60                                                  | 1568.16                                             | 1.017                            | .925                                                                   |
| 31.0                                                   | 7.2                                                  | 30.25                                  | 22.94                                                  | 526.24                                              | .589                             | — .008                                                                 |
| 22.6                                                   | — 8.4                                                | — 27.10                                | — 34.41                                                | 1184.05                                             | — .885                           | — 1.320                                                                |
| 18.2                                                   | — 4.4                                                | — 19.47                                | — 26.78                                                | 717.17                                              | — .688                           | — .474                                                                 |
| 10.9                                                   | — 7.3                                                | — 40.11                                | — 47.42                                                | 2248.66                                             | — 1.218                          | — .446                                                                 |
| 19.6                                                   | 8.7                                                  | 79.82                                  | 72.51                                                  | 5257.70                                             | 1.863                            | — 2.962                                                                |
| 18.0                                                   | — 1.6                                                | — 8.16                                 | — 15.47                                                | 239.32                                              | — .397                           | — .132                                                                 |
| —                                                      | —                                                    | —                                      | —                                                      | 18173.68                                            | —                                | — 6.630                                                                |

\* Standard deviation of y (*S.D.y*) = 38.92

$$r = \frac{\Sigma MSDx.MSDy}{n} = \frac{-6.630}{12} = -.553$$

$$\text{or } r = \frac{\Sigma xy}{n} = \frac{-6.630}{12} = -.553$$

tibles of the standard deviation. The principal advantage in modifying the first difference method in this way is that the magnitudes of first difference values are reduced. Table LXXVI will serve to illustrate the calculations involved in this method of correlation.

In column 2 are the statistics of production. Column 3 shows the differences between production of any one

<sup>1</sup> Based on data in Table LXXI, page 198.

year and production of the immediately preceding year. The percentage changes in production are recorded in column 4. These percentage changes are calculated by dividing the first differences of column 3 by the production of the respective preceding year. The mean of column 4 is obtained by summing the percentages algebraically and dividing by the total number of items, or percentage changes. In column 5 we have the deviations of the percentage changes in column 4 from their mean. The squares of these deviations are recorded in column 6. It is from the summation of column 6 that the standard deviation of the percentage changes of first differences in production is calculated. The multiples of standard deviation in column 7 are obtained by dividing the deviations of percentage changes in column 5 by the standard deviation. These multiples, therefore, as we have noted elsewhere in connection with a similar calculation in Table LXXIV, are merely the deviations in terms of the standard deviation.

The values in columns 9, 10, 11, 12, and 13 are obtained from the price data in column 8 by the method that has just been outlined for production in columns 3 to 7.

After the multiples of standard deviation are calculated, corresponding pairs are multiplied together, the products recorded in column 14, and then summated algebraically. The summation of this column is divided by the number of pairs of multiples to obtain the coefficient of correlation. Thus we have

$$\text{Coefficient of correlation } (r) = \frac{-6.630}{12} = -0.553.$$

For convenience the formula is ordinarily written as

$r = \frac{\Sigma xy}{n}$ , in which  $r$  is the coefficient of correlation,  $\Sigma xy$ ,

the summation of the products of corresponding pairs of multiples of standard deviation, and  $n$ , the number of pairs of items.

Like the first difference method, the percentage change of first difference method assumes each item to be a base. It goes farther than this, however, and converts the deviations of percentage changes into terms of multiples of the standard deviation, thereby removing practically all of the trend, if there is any.

It will be observed that the coefficients of correlation calculated by the two methods in Tables LXXV and LXXVI are not of the same magnitude, even though the same data are used. The principal reason for this is because the deviations of percentage changes from the mean of percentage changes do not always carry the same signs as the deviations themselves. In column 4 of Table LXXVI, for example, it will be seen that for 1923 the production of cotton was 3.95 per cent. greater than in 1922. This percentage change, however, was .28 per cent. less than the mean of all the percentage changes. It may be seen also, in Table LXXV, that for 1920 the first difference in production, column 2, is a plus quantity, while the corresponding first difference in price for 1920, column 5, is a minus quantity. Columns 3 and 6, though, show the deviations from the mean to be a minus quantity in both instances. Columns 5 and 11 in Table LXXVI show that for 1920 the deviation of the percentage change from the mean is a plus quantity in respect to production and a minus quantity for price. The principle underlying these calculations will be more readily understood after

the student has worked out by both methods a number of coefficients of correlation between different types of data.

#### PROBABLE ERROR OF THE COEFFICIENT OF CORRELATION

The calculation of the probable error of the coefficient of correlation has already been illustrated and discussed in Chapter IX. It will be recalled that the meaning of the probable error as applied to the coefficient of correlation is that one-half of any coefficients calculated from random selections are as likely to be larger than the first coefficient to the extent of the probable error as they are to be smaller to an equal extent or amount, the assumption being that the first coefficient was calculated from a random selection. This principle does not apply if the first sample is specially selected, while samples subsequently chosen are selected at random.

Probable errors of coefficients of correlation calculated from time series of economic data, however, do not have the usual interpretation that is placed upon probability. Any historical period selected is, as will be readily appreciated, a special period, with its own characteristics, often different from those of other periods. For this reason, therefore, such a series cannot be considered as constituting a random selection, since the individual items are not chosen independently. Consequently, the probable error of a coefficient of correlation calculated from time series does not indicate, as one might ordinarily conclude, that if a coefficient of correlation is calculated for other periods the chances are equal that these coefficients will fall within the range of the coefficient for the first period plus or minus the probable error. The probable error, then, of the coefficient of correlation between time series has little or no practical significance.



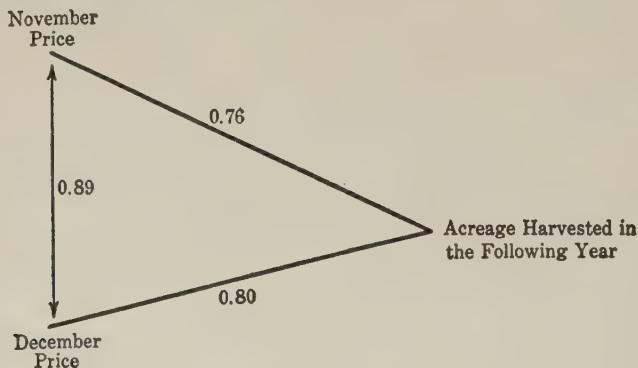
## MULTIPLE CORRELATION

Multiple correlation, as so obviously implied, is a measure of the combined effect of two or more independent variables upon one dependent variable. If it is found, for example, that the simple coefficient of correlation between rainfall during a certain period and yield of corn is less than 1, it is clearly indicated that some other factor or factors must be taken into account if we wish to measure the effect of all the independents upon yield. The significance of the coefficient of multiple correlation, then, lies in the fact that it is a precise numerical expression of the extent to which one dependent variable is related to or influenced by the joint or total effect of two or more causal factors. It may be observed further, let us say, that under certain circumstances the vegetative growth of plants is affected very favorably by a high percentage of sunshine and less favorably by heavy rainfall. A measurement of the relationship between growth and the joint or multiple effect of these two weather factors would be multiple correlation.<sup>1</sup>

Calculation of the coefficient of multiple correlation is rather simple. Let us illustrate by determining the multiple coefficient of correlation between the average spot quotations for middling upland cotton at New York and New Orleans for November and December and the acreage of cotton harvested in the subsequent crop year. Perhaps a simple diagram will make it easier to understand the method of calculating the multiple coefficient.

<sup>1</sup> An indication as to which of two independents affects the dependent to the greater extent may be obtained by calculating the product of their standard deviations. If this product approximates the magnitude of the standard deviation of the dependent, it is a fairly safe deduction to make that the independent with the highest coefficient of variation (standard deviation expressed as a percentage of the mean) exerts the greater effect upon the dependent.

From this diagram we may construct the equation for the calculation of the coefficient of multiple cor-



relation. The following form of equation is quite convenient:

$$RND.a = \sqrt{rNa.pNa + rDa.pDa}$$

In this equation the symbolic expressions have the following meanings:

$R$  = the coefficient of multiple correlation

$N$  = the November price

$D$  = the December price

$a$  = acreage harvested in the subsequent year

$RND.a$  = the coefficient of multiple correlation between  
November and December price and acreage

$rNa$  = the coefficient of correlation between November price and acreage

$pNa$  = the path coefficient of correlation between November price and acreage

$rDa$  = the coefficient of correlation between December price and acreage

$pDa$  = the path coefficient of correlation between December price and acreage.

By this equation the coefficient of correlation between November price and acreage is multiplied by the measure of net relationship, or path coefficient, between November price and acreage. This product is added to the product obtained by multiplying the coefficient of correlation between December price and acreage by the path coefficient of December price and acreage. Then the square root is extracted to obtain the coefficient of multiple correlation. This explanation implies that multiple correlation is exceedingly simple, which is true in many instances. Much of the failure on the part of students to grasp the technique of multiple correlation is due, perhaps, to cumbersome and complicated equations that are presented to them. The method outlined here has for its purpose the elimination of these complicated formulas and the substitution of simple equations that can be readily solved in many problems without a mastery of algebra.

By the method of percentage change of first differences it is found that the simple coefficient of correlation between deflated<sup>1</sup> average spot quotations for middling upland cotton for November at New York and New Orleans and acreage of cotton harvested the following year was .76 for the period from 1906 to 1923, while the coefficient between deflated average December prices at New York and New Orleans and subsequently harvested acreage of the next year for the same period is .80. The coefficient of correlation between November and December prices is seen to be .89.

Our first problem is to determine the net effect of both November and December price upon acreage. These measures of net effect are really the path coefficients. To

<sup>1</sup> By deflated prices in this instance is meant the original prices divided by index numbers of wholesale prices of the Bureau of Labor Statistics.

facilitate our calculation we will let  $D$  equal the net effect of December price and  $N$  equal the net effect of November price. Solving first for  $D$  we have:

$$\begin{array}{rcl} 1n + .89D & = & .76 \\ .89n + 1D & = & .80 \\ \hline .208D & = & .124 \quad (.208 \text{ obtained by multiplying } .89 \text{ by } \\ & & .89 \text{ and subtracting from } 1) \\ D & = & .596 \quad (.124 \text{ obtained by multiplying } .76 \text{ by } \\ & & .89 \text{ and subtracting from } .80) \end{array}$$

Solving for  $N$  we have:

$$\begin{array}{rcl} N + .89D & = & .76 \\ N + .53 & = & .76 \quad (.53 \text{ obtained by multiplying } .596 \text{ by } .89) \\ \hline N & = & .230 \quad (.230 \text{ obtained by subtracting } .53 \text{ from } \\ & & .76) \end{array}$$

The values of .596 for  $D$  and .230 for  $N$  are path coefficients, or measures of net relationship of prices for these months along the lines from November price to acreage and from December price to acreage.

Continuing the calculation we multiply the path coefficients by the simple coefficients, which gives

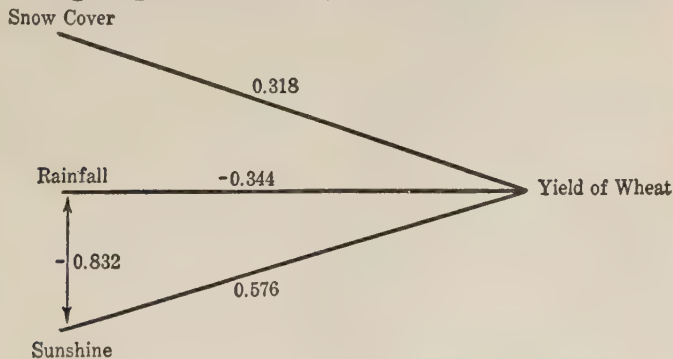
$$\begin{array}{rcl} .76 \times .230 & = & .175 \\ .80 \times .596 & = & .477 \\ \hline R^2ND.a & = & .652 \end{array}$$

The summation of .652 (.175 plus .477) is the degree of determination or the square of the coefficient of multiple correlation between November and December cotton prices and acreage of cotton harvested from the subsequent years' plantings. In other words, the value of .652 may be looked upon as the square of fluctuations in effect on cause. To obtain the coefficient of correlation we merely extract the square root of .652. We have, therefore,

$$RND.a = \sqrt{.652} = .807.$$

It will be observed that the coefficient of multiple correlation is expressed as  $R$ . As in the case of the simple coefficient of correlation, the multiple correlation coefficient expresses a measurement of the degree of relationship between the dependent and independent variables. No sign is attached because the measure as calculated is for a combination of relationships, some of which may be positive, whereas others are negative.

For further illustrating the calculation of the multiple coefficient of correlation we will take three independent variables and one dependent. The three independents are snow cover in late winter, sunshine in May, and rainfall in May at College Park, Maryland, and the dependent is the yield of wheat on the college farm at College Park. The following diagram will clarify the calculations involved.



The following symbolic expressions will be used in the calculation:

$S$  = snow cover

$p$  = rainfall

$s$  = sunshine

$y$  = yield.

For convenience we may construct the equation of  
 $RSps.y = \sqrt{rSy.pSy + rpy.ppy + rsy.psy}$ , in which

$R$  = the coefficient of multiple correlation

$S$  = snow cover in late winter

$p$  = rainfall in May

$s$  = sunshine in May

$y$  = yield of wheat

$RSps.y$  = the coefficient of multiple correlation between  
 the three independent variables, snow cover  
 in late winter, rainfall in May, and sunshine  
 in May, and the dependent variable, yield  
 of wheat

$rSy$  = the coefficient of correlation between snow  
 cover and yield of wheat

$pSy$  = the path coefficient between snow cover and  
 yield

$rpy$  = the coefficient of correlation between rainfall  
 and yield

$ppy$  = the path coefficient between rainfall and yield

$rsy$  = the coefficient of correlation between sun-  
 shine and yield

$psy$  = the path coefficient between sunshine and yield.

The simplicity of the calculations involved by this equation may be shown by outlining the procedure used in solving. This is as follows:

1. Multiply the coefficient of correlation between snow cover and yield by the path coefficient between snow cover and yield (.318 times .318).
2. Multiply the coefficient of correlation between rainfall and yield by the path coefficient between rainfall and yield (— .344 times .440).



3. Multiply the coefficient of correlation between sunshine and yield by the path coefficient between sunshine and yield (.576 times .942).
4. Summate the products in 1, 2, and 3. The product in 2 (−.344 times .440) is a minus quantity, so that in summing it will be subtracted from the total of the other two products.
5. Extract the square root of the summation in 4. The result is the coefficient of multiple correlation.

It will be observed from the diagram that no inter-relationship is calculated between snow cover in late winter and rainfall in May. This is because that over a period of years it is reasonable to suppose there will be no causal relation between these two factors. In the calculation of the multiple coefficient of correlation, therefore, the effect of snow cover is held constant, with only the multiple effect of rainfall and sunshine being measured.

Proceeding with the calculation we have:<sup>1</sup>

$$\begin{array}{rcl}
 p - .832s = - .344 & (.308 \text{ obtained by multiplying} \\
 - .832p + 1s = .576 & .832 \text{ by } .832 \text{ and subtracting} \\
 \hline & \text{from 1)}
 \end{array}$$

$$.308s = .290$$

$$\begin{array}{rcl}
 s = .942 & (.290 \text{ obtained by multiplying} \\
 & - .344 \text{ by } - .832 \text{ and subtracting} \\
 & \text{from } .576)
 \end{array}$$

$$\begin{array}{rcl}
 - .784p = - .344 & (- .784 \text{ obtained by multiply-} \\
 & \text{ing } - .832 \text{ by } .942)
 \end{array}$$

$$\begin{array}{rcl}
 p = .440 & (.440 \text{ obtained by dividing} \\
 & - .344 \text{ by } - .784)
 \end{array}$$

<sup>1</sup> As in the preceding calculation of the coefficient of multiple correlation, the first step is to determine the path coefficients, or the net effect of each independent variable upon the dependent. This is merely a matter of solving simple normal equations.

The values of .942 and .440 are path coefficients for sunshine and rainfall respectively. The path coefficient between snow cover and yield is taken as .318, the same as the simple coefficient of correlation, since, as has been stated, it can hardly be supposed that over a period of years there is any causal relation between snow cover in late winter and rainfall in May. The value of .318, then, for our purpose, is the path coefficient between snow cover and yield. Multiplying the path coefficients by the simple coefficients of correlation we have:

$$\begin{array}{rcl} .318 \text{ times } .318 & = & .1011 \\ -.344 \text{ times } .440 & = & -.1514 \\ .576 \text{ times } .942 & = & .5426 \end{array}$$


---

$$R^2 \text{ (coefficient of determination)} = .4923$$

$$R = \sqrt{.4923} = .702.^1$$

The value .702 is the coefficient of multiple correlation between snow cover in late winter, rainfall in May, sunshine in May, and yield of wheat per acre, with snow cover in late winter held constant.

In any other calculations of coefficients of multiple correlation the same method may be followed, the simple and path coefficients being added to the equation already formulated. After a few problems have been worked out the student will realize how exceedingly simple it is to calculate the coefficient of multiple correlation.

### PARTIAL CORRELATION

Sometimes it may be desired to measure the relationship between one independent and the dependent with the effect of other independent variables held constant or eliminated. A coefficient of correlation thus calculated

<sup>1</sup> Based on data in Table LXXX, page 257, Chapter XIII.

is in effect the net relationship between the dependent variable and a single independent variable, since by this method the effect of other independents taken into account is considered as remaining unchanged from time to time. Holding of an independent variable constant has already been illustrated in the preceding calculation of the coefficient of multiple correlation. To illustrate the method of partial correlation further, we will calculate the net relationship between supply of cotton and price of cotton in the United States with the effect of exports of cotton held constant. To simplify the calculation we may construct the following equation:

$$rsp.e = \frac{rs.p - (rs.e \text{ times } re.p)}{\sqrt{(1-rs.e^2)(1-re.p^2)}}$$

The following are the meanings of the symbolic expressions in the equation.<sup>1</sup>

*rsp.e* = the coefficient of partial correlation between supply and price, with the effect of exports held constant

*rs.p* = the simple coefficient of correlation between supply and price

*re.p* = the simple coefficient of correlation between exports and price

*rs.e* = the simple coefficient of correlation between supply and exports

*rs.e*<sup>2</sup> = the square of the coefficient of correlation between supply and exports

*re.p*<sup>2</sup> = the square of the coefficient of correlation between exports and price.

<sup>1</sup> Any other symbols, with the exception of *r*, could just as well be used. The letter *r* is always used to denote simple correlation.

By the method of percentage change of first differences for a period of years the following expressions or measurements of relationship were found:

$$r \text{ between supply and price} = -.908$$

$$r \text{ between exports and price} = -.754$$

$$r \text{ between supply and exports} = .663$$

Substituting these values in the equation we have:

$$\begin{aligned} rsp.e &= \frac{-.908 - (.663 \text{ times } -.754)}{\sqrt{(1 - .663^2)(1 - .754^2)}} = \frac{-.408}{\sqrt{.560 \times .431}} \\ &= \frac{-.408}{.491} = -.83. \end{aligned}$$

The details of the preceding calculation consist first of subtracting from the coefficient of correlation between supply and price,  $-.908$ , the product of the coefficients of correlation between supply and exports and exports and price. These coefficients are  $.663$  for supply and exports and  $-.754$  for exports and price. Multiplying  $.663$  by  $-.754$  gives a product of  $-.500$ . Removing the brackets in the equation changes the minus sign belonging to  $.500$  to a plus, since the sign preceding the brackets is a minus, so that the product of  $-.500$  becomes  $+.500$ . Subtracting  $.500$  from  $-.908$  leaves  $-.408$ . If the sign preceding the brackets had been a plus, the sign inside the brackets would not have changed when the brackets were removed. This is a principle to be observed in all similar calculations.

The next step is to square the coefficient of correlation between supply and exports,  $.663$ , and then subtract from 1, which gives  $.560$ . Then we square the coefficient of correlation between exports and price,  $-.754$ , and subtract from 1, obtaining a result of  $.431$ . Multiplying  $.560$

by .431 and extracting the square root of the product gives .491, which when divided into  $-.408$  gives  $-.83$ , the coefficient of partial correlation.

The significance of this coefficient of  $-.83$  is that if it were possible to select a sufficient number of years in which the exports were essentially the same, but which were otherwise normal, the correlation between supply and price of cotton for such years would tend to be around  $-.83$ .

Calculation of the coefficient of partial correlation with one independent held constant furnishes a basis for similar calculations when the effects of several independent factors are to be held constant. All that is necessary is to modify the equation so as to include the additional simple coefficients of correlation in the numerator and denominator.

### RANK-ORDER CORRELATION

Although rank-order correlation is not in common use, it may be well to have an understanding of the method involved in calculating the coefficient. If, for example, we know the relative ranks of items in each of two series but do not know the actual size of the items themselves, how are we to draw a conclusion as to the extent of relationship between corresponding pairs of items in the two series? The calculations on which to base a conclusion in such instances will be clarified by referring to Table LXXVII. Here are shown, in columns 2 and 3, the relative ranks of cotton production and cotton prices in the United States for the years 1923 to 1928. These ranks alone give us a very definite idea of relationship between production and price. In 1926, for instance, production was higher than in any of the other years,

whereas price was lower. On the other extreme, production was lowest in 1923, with price higher than in any other year. For a definite measure of this relationship the formula,

$$p \text{ (rank-order correlation)} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \quad ^1$$

has been found very convenient, in which  $\sum d^2$  is the summation of the squares of deviations in ranks of the two series, and  $n$  the number of pairs of ranks. Column 4 of the table shows the deviations in ranks, or the differences between columns 2 and 3, while the squares of these deviations are recorded in column 5. Substituting the proper values we have:

$$\begin{aligned} p \text{ (rank-order correlation)} &= 1 - \frac{6\sum d^2}{n(n^2 - 1)} \\ &= 1 - \frac{396}{210} = -.886. \end{aligned}$$

The numerator, 396, is 6 times 66, the summation of column 5, while the denominator, 210, is obtained by squaring the number of pairs of ranks, which is 6, subtracting 1, which gives 35, and then multiplying 35 by 6, which is  $n$ . The quotient obtained by dividing the product of 35 and 6 (which is 210) into 396 is 1.886. Subtracting 1.886 from 1 leaves  $-.886$ , the coefficient of rank-order correlation. This coefficient indicates that as production of cotton increases or decreases there is a change in price in the opposite direction. In other words, it indicates that there is inverse or negative relationship between production and price of cotton, which ordinarily is to be expected.

<sup>1</sup> Spearman's rank-order formula in modified form. The details of this formula and its derivation may be reviewed in *Statistical Methods*, by T. L. Kelley, pages 191 to 193.



TABLE LXXVII. RELATIVE RANK OF PRODUCTION AND PRICE OF COTTON IN THE UNITED STATES, 1923-1928 <sup>1</sup>

| 1     | 2                  | 3                                           | 4                                                      | 5                            |
|-------|--------------------|---------------------------------------------|--------------------------------------------------------|------------------------------|
| YEAR  | RANK IN PRODUCTION | RANK IN PRICE RECEIVED BY PRODUCERS, DEC. 1 | DEVIATION IN RANK (Difference between Columns 2 and 3) | SQUARE OF DEVIATIONS IN RANK |
| 1923  | 6                  | 1                                           | 5                                                      | 25                           |
| 1924  | 4                  | 2                                           | 2                                                      | 4                            |
| 1925  | 2                  | 4                                           | 2                                                      | 4                            |
| 1926  | 1                  | 6                                           | 5                                                      | 25                           |
| 1927  | 5                  | 3                                           | 2                                                      | 4                            |
| 1928  | 3                  | 5                                           | 2                                                      | 4                            |
| Total | —                  | —                                           | —                                                      | 66                           |

## CORRELATION INDEX

When the relationship between corresponding pairs of variables in two series is nonlinear,<sup>2</sup> or, in those instances when it cannot be represented by a straight line, some measure other than the coefficient of correlation may be used to express this relationship. For such purpose the correlation index has been found to be convenient, having values that range from 0 to 1. The calculation of this measure is simple. Let us turn back to Chapter XI for illustrative data. In Table LXX are the statistics of acreage and yield of corn per acre on twenty-seven farms in Georgia and Indiana in 1928. The data are graphically presented in Chart 37. As will be readily seen, the relationship between acreage of corn on individual farms and yield per acre is nonlinear, since it cannot properly be expressed or represented by a straight line. It is on relationships of this type, therefore, that the correlation index is based.

<sup>1</sup> *Yearbook, United States Department of Agriculture, 1928*, page 837, Table 250.

<sup>2</sup> Nonlinear relationship is sometimes referred to as curvilinear relationship.

The details of calculating the correlation index are very readily understood. The first step is to determine the deviations of the items in the dependent series from the line of best fit. These deviations are then squared, the products summated, divided by the number of deviations, and the square root extracted. The result thus obtained is the standard deviation, a measure of dispersion from trend or line of best fit. Next the standard deviation of actual values is calculated. The following formula may be used in simplifying the calculations:

$$ci = \sqrt{1 - \frac{Sy^2}{SDy^2}}, \text{ in which}$$

$ci$  = the index of correlation

$Sy^2$  = the square of the standard deviation of  $y$  values  
calculated from deviations from trend

$SDy^2$  = the square of the standard deviation of  $y$  values  
calculated from deviations from the arithmetic  
mean of actual values.

It will be seen from this formula that by dividing  $Sy^2$  by  $SDy^2$  and then subtracting from 1, the square of the correlation index is obtained, so that the correlation index itself is then obtained by merely extracting the square root of the resulting value. The principal significance of the index of correlation is that it measures the degree of dispersion about a curve which has been fitted to the data either by mathematical or by free-hand methods. It is obvious that the measure thus calculated, like the correlation ratio, carries no sign.

### CORRELATION RATIO

Another measure of correlation between series in which the relationship between pairs of items is nonlinear is the

correlation ratio. This measure is based on the deviations of the means of columns from the mean of the entire distribution. On the x axis, for example, we may have the following classes: 1-5, 6-10, 11-15, 16-20, and 21-25. On the y axis the values in the 1-5 class, we will say, range from 1 to 12, in the 6-10 class they range from 10 to 21, from 17 to 29 in the 11-15 class, from 26 to 37 in the 16-20 class, and from 34 to 43 in the 21-25 class. The first step is to calculate the mean of each class, then the mean of all the items in the classes taken as a whole. Next the deviations of the means of the several columns or classes from the mean of all the items are determined, the deviations squared, multiplied by the number of items in the columns or classes, the products summated, and then divided by the number of items. The result thus obtained is the mean of the squares of deviations. The square root of this mean of squares is extracted to obtain the standard deviation of the means of columns ( $SDmc$ ), which is divided into the standard deviation of the entire array or series ( $SDty$ ) to obtain the correlation ratio.

The measure thus calculated falls within the range of 0 and 1, and, like the index of correlation, carries no sign. It is simply a measure of the extent to which a line passing through the means of columns fits the data.

The correlation ratio may be defined also as the measure of similarity between the deviations from the means of individual arrays or columns and the deviations from the mean of the entire data. If the items in the columns centered on the mean of the columns, the standard deviation as calculated from the deviations of all the items in all the columns would be the same as the standard deviation calculated from the deviations of the means of individual

columns from the mean of all the columns, or rather the mean of all the items in all the columns. To the extent that the items in the columns fluctuate from their respective means, the standard deviation calculated from the fluctuations of the means of columns from the mean of all the items will be smaller than the standard deviation based on deviations of individual items from the mean of the entire group. This is the reason that in determining the correlation ratio the standard deviation of the items considered in their entirety is divided into the standard deviation calculated from the deviations of means of columns from the mean of the entire data.

For convenience the following formula may be evolved:

$$cr \text{ (correlation ratio)} = \frac{SDmc}{SDty},$$

in which

$SDmc$  = the standard deviation of the means of columns as calculated from the square of the deviations of each mean from the mean of all the items and the number of items in the respective columns

$SDty$  = the standard deviation of the total group of items.

Generally the correlation ratio shows a higher degree of relationship than the correlation index, for which reason it may give erroneous impressions unless properly interpreted. The important significance of this measure is that it indicates the degree of dispersion about a line connecting the means of the columns of a correlation table. Its application is often most desirable when there is a very large number of items involved in each class or group.

## CORRELATION TABLE

If it is desired to show the direction of correlation without constructing a chart or expressing the relationship in numerical terms the data may be arranged in the form of a frequency table to indicate the association of corresponding pairs of variables. We have already seen in Chapter XI how to construct lines and curves showing linear and nonlinear relationships. Using the data in Table LXX in the preceding chapter on which Chart 37 is based we may now construct another table that shows in a general way the relationship between two series of items.

FARMS TABLE LXXVIII. ACREAGE AND YIELD OF CORN ON SELECTED FARMS IN GEORGIA AND INDIANA IN 1928 <sup>1</sup>

(Correlation Table)

| ACREAGE | NUMBER OF FARMS PRODUCING WITHIN THE LIMITS OF SPECIFIED YIELDS PER ACRE IN BUSHELS |       |       |       |       |       |       | TOTAL |
|---------|-------------------------------------------------------------------------------------|-------|-------|-------|-------|-------|-------|-------|
|         | 9-14                                                                                | 15-20 | 21-26 | 27-32 | 33-38 | 39-44 | 45-50 |       |
| Total   | 5                                                                                   | 2     | 2     | 2     | 6     | 6     | 4     | 27    |
| 1-5     | 5                                                                                   | 1     | 1     |       |       |       |       | 7     |
| 6-10    |                                                                                     | 1     |       | 1     |       |       |       | 2     |
| 11-15   |                                                                                     |       | 1     | 1     | 3     | 1     |       | 6     |
| 16-20   |                                                                                     |       |       |       |       | 2     | 1     | 3     |
| 21-25   |                                                                                     |       |       |       |       | 2     | 1     | 3     |
| 26-30   |                                                                                     |       |       |       | 3     | 1     | 2     | 6     |

This table is so constructed that the reader sees at a glance the tendency for high yields per acre to be directly associated with the larger acreages.

<sup>1</sup> Based on data in Table LXX, page 192.

### CALCULATION OF THE COEFFICIENT OF CORRELATION FROM A FREQUENCY DISTRIBUTION

When associated items have been tabulated or arranged in the form of a frequency table the coefficient of correlation may be readily calculated directly without the necessity of any rearrangement of the items. The method of constructing a frequency correlation table which will serve for making the necessary extensions involved in calculating the coefficient of correlation is illustrated in Table LXXIX.

The calculations involved in this table are very simple. It will be seen from the *f* column that there are 20 double frequencies, or 20 pairs of frequencies. That is to say, there are 20 production items and 20 price items. These quantities or summations are given in the "Total" columns. The columns showing the deviations in class interval units from the assumed means are headed *d'*. In the *x* series, production, the midpoint of the 352-402 class is taken as the assumed mean. This midpoint is 377. The next step now is to show the extent to which all of the other midpoints of classes deviate from the assumed mean. The assumed mean itself, of course, shows a deviation of 0. Proceeding, we see that the midpoint of the 301-351 class, 326, deviates -1 from the assumed mean of production, while the midpoint of the 250-300 class deviates -2 from the assumed mean. On the other side of the assumed mean it will be seen that the midpoints of the two classes deviate +1 and +2, respectively, from the point of origin, which is the midpoint of the 352-402 class.

On the *y* axis the midpoint of the 111-141 class is selected as the assumed mean. This midpoint is 126. As is true with production, the deviations in class interval units of



TABLE LXXIX. RELATION BETWEEN PRODUCTION AND PRICE OF POTATOES IN THE UNITED STATES, 1909-1928 <sup>1</sup>

(Calculation of the Coefficient of Correlation from a Frequency Distribution)

|                                                           |                 | x PRODUCTION IN MILLIONS OF BUSHELS |    |     |                    |         |               |               |                 |               |       | (x'y') |
|-----------------------------------------------------------|-----------------|-------------------------------------|----|-----|--------------------|---------|---------------|---------------|-----------------|---------------|-------|--------|
| y Price in cents per bushel received by producers, Dec. 1 | Class intervals |                                     |    |     |                    | 250-300 | 301-351       | 352-402       | 403-453         | 454-504       | Total |        |
|                                                           | Mid-points      |                                     |    |     |                    | 275     | 326           | 377           | 428             | 479           |       |        |
|                                                           |                 | f                                   |    |     |                    | 2       | 4             | 4             | 9               | 1             | 20    |        |
|                                                           |                 |                                     | d' |     |                    | -2      | -1            | 0             | + 1             | +2            | 0     |        |
|                                                           |                 |                                     |    | fd' |                    | -4      | -4            | 0             | + 9             | +2            | + 3   |        |
|                                                           |                 |                                     |    |     | f(d') <sup>2</sup> | 8       | 4             | 0             | 9               | 4             | 25    |        |
|                                                           | 49-79           | 64                                  | 10 | -2  | -20                | 40      | +2<br>2<br>+4 | 0<br>2<br>0   | - 2<br>5<br>-10 | -4<br>1<br>-4 | 10    | -10    |
|                                                           | 80-110          | 95                                  | 3  | -1  | - 3                | 3       | +2<br>1<br>+2 | 0<br>1<br>0   | - 1<br>1<br>- 1 |               | 3     | + 1    |
|                                                           | 111-141         | 126                                 | 4  | 0   | 0                  | 0       |               | 0<br>1<br>0   | 0<br>3<br>0     |               | 4     | 0      |
|                                                           | 142-172         | 157                                 | 2  | +1  | + 2                | 2       | -2<br>1<br>-2 | -1<br>1<br>-1 |                 |               | 2     | - 3    |
|                                                           | 173-203         | 188                                 | 1  | +2  | + 2                | 4       |               | -2<br>1<br>-2 |                 |               | 1     | - 2    |
|                                                           | Total           |                                     | 20 | 0   | -19                | 49      |               |               |                 |               | 20    | -14    |

<sup>1</sup>Basic data tabulated from *Yearbook, United States Department of Agriculture, 1928*, page 806, Table 202.

midpoints of classes below the assumed mean carry the minus sign, whereas the deviations above the assumed mean are plus deviations. Thus it is that the midpoint of the 80-110 class, 95, deviates  $-1$  from 126, the assumed mean, while the midpoint of the 49-79 class, 64, deviates  $-2$ . The midpoints of the 142-172 and 173-203 classes, lying above the mean, deviate  $+1$  and  $+2$ , respectively, from 126, the point of origin on the  $y$  axis.

The values in the  $fd'$  columns, constituting the total deviations from the assumed mean, are the products of the frequencies and the deviations of midpoints from the assumed means, while  $f(d')^2$  is the product of the frequencies and the square of the deviations. These products, that is, the  $f(d')^2$  values, are, in fact, the squares of the deviations from the assumed means. A small summation of  $fd'$  and of  $f(d')^2$  indicates that the assumed mean approximates the true mean.

Perhaps by calculating the true mean of each of the two series arranged in Table LXXIX the fundamental principles involved in the other calculations will be more readily comprehended. Let us take up the  $x$ , or production, series first. We have already seen that the assumed mean of this series is 377. But this does not coincide with the true mean of production, since the summation of the  $fd'$  column is not zero. We see, therefore, that with an  $fd'$  summation of  $+3$  the true mean is not only different from the assumed mean, but it is also greater than the assumed mean of 377. The problem is to merely determine the correction factor. The average of  $+3$ , the summation of the  $fd'$  column, is obtained by dividing it by 20, the total number of frequencies. Thus, we have

$$cmx \text{ (correction factor for the mean of } x) = \frac{+3}{20} = +.15,$$

which is in terms of class interval units because the summation of  $fd'$  is in class interval units. This value of  $+.15$  shows that the true mean is  $.15$  of a class interval greater than the assumed mean. To convert  $+.15$  to terms of actual values we simply multiply it by the size of the class interval, which is 51. Therefore, we have  $+.15$  times  $51 = +7.65$ , which indicates that the true mean is 7.65 greater than the assumed mean of 377. Ignoring the fractional part of  $+7.65$  as such, and expressing the correction factor in whole numbers, we have  $+8$ . This leads to the following calculation:

$$\text{True mean of production} = 377 + 8 = 385.^1$$

In the consideration of price, 126 was selected as the assumed mean. The summation of  $fd'$  is seen to be  $-19$ , which not only shows that the true mean is different from the assumed mean of 126, but that the true mean is less than the assumed mean, since the summation of the deviations  $-19$ , is a negative quantity. As was done in calculating the correction factor for the mean of the  $x$  series, the correction factor for the mean of  $y$  is obtained by dividing the  $fd'$  summation by 20, the total number of frequencies. Thus, we have

$$cm_y \text{ (correction factor for the mean of } y) = \frac{-19}{20} = -.95$$

in class interval units. That is to say, the true mean is less than the assumed mean to the extent of  $.95$  of a class

<sup>1</sup> It will be understood, of course, that the so-called "true mean" calculated by this method may not actually be the same as the mean obtained by summing all the individual items and dividing by the total number of items. This is because in Table LXXIX the calculation of the  $fd'$  summation is based on the assumption that the individual frequencies in each class deviate to an equal extent on either side of the midpoint, which would seldom be expected to be true. It is for this reason, as has been indicated elsewhere in this text, that certain limitations must be placed upon the significance of a mean calculated from a frequency distribution when the midpoints are used as means.

interval. To convert  $-.95$  to terms of actual values it is multiplied by the size of the class interval. This class interval is 31, as will be noted. We have, therefore,  $-.95$  times 31 =  $-29.45$ , showing that the true mean is 29.45 less than the assumed mean. The calculation, then, becomes

$$\text{True mean of price} = 126 - 29.45 = 96.55.^1$$

Proceeding to the calculation of the coefficient of correlation, we will first determine the value of  $\Sigma(x'y')$ . A brief explanation will show how the individual products making up this summation are obtained. Let us illustrate by referring to the 49-79 class on the y axis. We see in the f column that there are 10 frequencies in this class. On the x axis it will be seen that 2 of these 10 frequencies belong to the 301-351 class, 2 to the 352-402 class, 5 to the 403-453 class, and 1 to the 454-504 class. The values above these x frequencies are the products of deviations from the assumed means. Hence, +2 above the frequency 2 in the y class 49-79 and in the x class 301-351 is the product of the x and y deviations from the assumed means. The deviation on the x axis is  $-1$  and on the y axis the deviation is  $-2$ . Multiplying  $-1$  by  $-2$  gives a product of  $+2$ . The lower product or value, 4, is obtained by multiplying  $+2$  by 2. All the other products are obtained in the same way, the individual products summated algebraically, and the summation recorded in the column headed  $\Sigma(x'y')$ . The total of the individual summations thus obtained is the value for  $\Sigma(x'y')$ . In Table LXXIX this total summation is  $-14$ .

The next step is to calculate the standard deviation of each of the two series. We will take the x series first.

<sup>1</sup> See footnote relating to calculation of the true mean of production, page 241.

In the  $f(d')^2$  column there is a summation of 25, which, as has been seen, is the total of the squares of the deviations from the assumed mean. The standard deviation, though, is based on the squares of deviations from the true mean. It becomes necessary, therefore, to make an adjustment in the  $f(d')^2$  summation of 25. It was found in calculating the mean of  $x$  that the correction factor for the true mean is  $+.15$ , in terms of class interval units. The square of  $.15$  is  $.02$ , which is added to the mean of 25, the  $f(d')^2$  summation, to obtain the mean of the squares of deviations from the true mean, which will be the square of the standard deviation of  $x$ . For convenience this formula may be used:

$$S.D.^2 = \frac{\Sigma f(d')^2}{N} + c^2 = \frac{25}{20} + .02 = 1.27,$$

in which  $c$  is the correction factor. The standard deviation is now obtained by the formula

$$S.D. = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{1.27} = 1.12.$$

This standard deviation is in terms of class interval units because the summation of the squares of deviations from the mean is in these terms. The standard deviation in actual units of production is obtained by simply multiplying 1.12 by the size of the class interval, which is 51. We have, therefore,

$$S.D. \text{ in actual units of production} = 1.12 \text{ times } 51 = 57.12.$$

The standard deviation of  $y$  is calculated in the same way as the standard deviation of  $x$ . The summation of the  $f(d')^2$  products is 49, which is the total of the squares of the deviations from the assumed mean of price. The mean of the squares of deviations from the true mean is

obtained by subtracting from the mean of 49, summation of  $f(d^1)^2$ , the square of the correction factor. Dividing 49 by 20, the number of frequencies, gives 2.45, the mean of the squares of deviations from the assumed mean. The correction factor, as we have seen, is  $-.95$ . Subtracting the square of  $-.95$  from 2.45 gives the mean of the squares of deviations from the true mean, or the square of the standard deviation of price. Accordingly, we have

$$S.D.^2 = \frac{\Sigma f(d')^2}{N} - c^2 = \frac{49}{20} - (-.95)^2 = 1.55,$$

in which  $c$  is the correction factor. The standard deviation is now obtained by means of the formula

$$S.D. = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{1.55} = 1.25.$$

The standard deviation thus calculated, like that of production, is in terms of class interval units. To convert it to actual values of original units of price is to merely multiply it by the class interval, which is 31. Thus, we have

$$S.D. \text{ in actual units of price} = 1.25 \text{ times } 31 = 38.75.$$

We must now determine the mean product. This is done by subtracting the product of the two correction factors as calculated in terms of class interval units from the mean of  $\Sigma(x'y')$ . We have already seen that the correction factor for the production series is  $+.15$ , whereas for price it is  $-.95$ . For convenience in calculating the mean product we have

$$p \text{ or } \Sigma(xy) = \frac{\Sigma(x'y')}{N} - cxcy,^1$$

<sup>1</sup> F. C. Mills, *Statistical Methods*, 1924, page 392.



in which  $cx$  and  $cy$  are correction factors in terms of class interval units. In the preceding equation,

$$\Sigma(x'y') = -14$$

$$cx = +.15$$

$$cy = -.95$$

$$cxcy = -.14$$

$$N = 20.$$

Therefore, substituting, we have

$$p = \frac{-14}{20} - (-.14) = -.7 + .14 = -.56$$

in class interval units.

The coefficient of correlation is now obtained by dividing the value of  $p$ ,  $-.56$ , by the product of the two standard deviations in terms of class interval units. For this calculation we have

$$r = \frac{p}{SDx.SDy} = \frac{-.56}{1.12 \times 1.25} = \frac{-.56}{1.39} = -.40$$

This is a rather low expression of correlation, owing partly to the fact that by this method of correlation the items or frequencies in the classes are assumed to be evenly distributed about the midpoints of classes. To the extent that this is not true, there will be the tendency for the numerical expression of correlation to be minimized. By referring to the original data it will be observed, for example, that both of the  $x$  frequencies in the 250–300 class are greater in size than the midpoint of the class.<sup>1</sup> Therefore, the deviation of  $-2$  does not reflect the actual class-interval deviations in those items. In other words, the deviations of the two items in the 250–300 class are

<sup>1</sup> These two figures are of the magnitudes 293,000,000 and 287,000,000, respectively. *Yearbook, United States Department of Agriculture, 1928*, page 806, Table 202.

both greater than two class intervals from the point of origin. This indicates what might possibly be expected in the other classes of  $x$  items, and also what condition may exist in the classes of  $y$  items. Another factor that should be considered is the relative size of the class intervals in the two series. Sometimes, of course, it is not possible to form any definite idea as to what the mathematical ratio between the magnitudes of class intervals in two series should be until the coefficient of correlation itself has been calculated. In this problem we have a class interval of 51 in the  $x$  series and an interval of 31 in the  $y$  series. It is quite probable that a much higher coefficient of correlation would be obtained by materially reducing the class interval in the  $y$  series. For the sake of simplicity in illustrating the calculation of the coefficient of correlation from a frequency table, the intervals of 51 and 31, respectively, have been selected so that there will be the same number of classes in each of the two series.

It is only when the number of frequencies is relatively large that the coefficient of correlation should be calculated from a frequency table, unless the proper ratio is provided for between the intervals of the  $x$  and  $y$  series. The data in Table LXXIX were selected merely to illustrate method of calculation, and not to convey an idea of the extent of relationship between production and price of potatoes.

#### TIME LAG

In discussing and illustrating the general principles and technique of correlation no mention was made of the fact that sometimes the fluctuations in one series are not accompanied by immediate fluctuations in another. That is to say, the influence of changes may not be evident

until after the lapse of several months, or even of a year or more. It would hardly be expected, for instance, that a sudden permanent rise in farm wages would be accompanied or followed by an immediate increase in the price of farm products, unless, of course, other factors were operating at the same time. In an analysis of cotton price movements it was found that spot quotations for middling upland cotton tended toward an increase or decrease twelve or thirteen months after an upward or downward trend in general business conditions as reflected by the prices of industrial stocks. This, then, represents a lag of twelve or thirteen months in cotton prices. Another analysis showed an increase or decrease in pig iron production to be followed about seven months later by an increase or decrease in spot prices of middling cotton. In calculating coefficients of correlation between series in which there is a lag it is a good plan to plot each one separately and then determine the approximate lag by shifting back and forth.

In regard to time lag the student must remember that it is only when changes in a series move in cycles or when they are permanent that sympathetic changes can be expected in another series.

Those who are interested in more advanced methods of correlation may consult the works of Doctors H. R. Tolley and M. J. B. Ezekiel, two of the most outstanding present-day statisticians and economists. Students of biology, as well as those pursuing courses in economics, will find Dr. Wm. B. Kemp, University of Maryland, College Park, Maryland, an eminent authority also.

## CHAPTER XIII

### REGRESSION

The meaning of regression has already been indicated in Chapters IV and XI. As applied to statistics it is, briefly stated, the tendency, when corresponding pairs of items or values are correlated, that is manifested by the values in either series to revert to their characteristic type rather than to necessarily exhibit the particular amount of variation or change that may be observed in the other series. The term is in reality a borrowing from the field of biology. It has been observed, for example, by geneticists that offspring do not inherit the full amount of certain characteristics or peculiarities of the parents, so that if both parents are an inch taller than the average height of individuals, the chances are that their offspring will be less than an inch above the average. Thus it is that the offspring are likely to revert or regress toward the type, which in this instance is the average height. It is from this tendency, therefore, that the term "regression" is derived.

To indicate the degree of relationship between corresponding pairs of items in two series the line of regression may be drawn. This is merely a line connecting the plotted ordinates of highest relationship between the corresponding pairs of items, so fitted that the sum of the squares of deviations is less than the sum of the squares of deviations from the calculated ordinates of any other similar

line of relationship. Furthermore, the values of  $y$ , the dependent variable, that are calculated from the regression equation bear a higher relationship to actual values of  $y$  than when calculated from any other like or similar type of relationship between  $x$  and  $y$  values. It is because of this, to the extent that both linear and nonlinear relationships can be used for predictive purposes, that the regression equation, or the equation used in fitting the line of regression, gives ordinates of the line of relationship that are the closest estimates of  $y$ , the dependent, that are most likely to be associated with corresponding values of  $x$ , the independent variable.

As calculated, the coefficient of regression, a component part of the equation used in deriving the ordinates of the regression line, is an expression indicating the extent to which a change in a given value in a series is associated with a like or opposite change in another series. It measures under certain conditions the amount of this change or effect that is most likely to occur. For example, if we were correlating pig iron production in the United States with prices of pig iron, with the former expressed in units of thousands of tons and the latter in units of one dollar per ton, and found the coefficient of regression to be  $-.65$ , after the coefficient of correlation had been calculated, the indications would be that for each change of one thousand tons in production above or below the average production there would be a change in price from the average in the opposite direction to the extent of  $.65$  of one dollar per ton. If production, say, in 1930 is found to be one unit, or 1,000 tons, below the average production for a certain period of years, the price per ton, on the basis of the relationship assumed, would be likely to increase only  $.65$  of one unit, or  $.65$  of one dollar. By this



illustration it is seen that the change in price, the dependent variable, is less than the change in production, the independent variable. It is in this sense, therefore, that price would show, on the average, a change of less magnitude than the change in production, regressing toward the mean of all the prices in the class or period. A review and further inquiry into the fundamental principles of regression will more fully clarify its meaning.

### SIMPLE REGRESSION

When we have measured the extent to which corresponding pairs of items in two series are, on the average, associated, or the extent to which changes in one independent series are accompanied or followed by similar or opposite changes in a dependent series, the result obtained is a numerical expression of simple regression. The name "simple regression" is derived from the fact that this expression of association, which may be termed also an expression of probability or estimate, is based on but one cause and effect relationship. For example, in calculating the ordinates of the regression line showing linear relationship between cotton production and cotton prices in Chart 36, Chapter XI, there is taken into account only the effect of production upon price. This regression line, then, is merely the line showing the highest average relationship between cotton production and cotton prices, with all other factors ignored that may have some relation to those prices. This is the simplest form of regression. There must, of course, be two items that are paired, or two series of corresponding pairs, as the case may be, before there is any basis for calculating a numerical expression of relationship.

In calculating the ordinates of the regression line by



the method illustrated in Chart 36, Chapter XI, the simple arithmetic mean of each of the two series must be determined, as well as the coefficient of correlation between corresponding pairs of items in the two series. The regression coefficient of  $y$  on  $x$  is itself obtained, as has been seen, from the formula  $b = r \frac{\sigma_y}{\sigma_x}$ , and the values of

$y$ , the dependent, most likely to be associated with given values of  $x$ , the independent, are obtained by employing the ordinary formula already explained for straight line relationship,  $y = a + bx$ , in which  $y$  is the ordinate of the straight line of average relationship,  $a$  the arithmetic average,  $b$  the coefficient of regression, and  $x$  the distance on the  $x$  axis from the point where the mean of the  $x$  items coincides with the mean of the  $y$  items. The standard deviation is calculated either by squaring the deviations from the mean, summing, dividing by the number of deviations, and extracting the square root, or by squaring the actual items, summing, dividing by the number of items, subtracting the square of the mean, and then extracting the square root.

An important generalization to make for the benefit of the beginner is that the coefficient of correlation and the standard deviations used in the formula for the coefficient of simple rectilinear regression should be associated. For instance, if the Pearsonian coefficient is used in obtaining the regression coefficient for straight line relationship, then the standard deviations calculated from the squares of the deviations of actual values from the mean of actual values should also be used, unless for some special reason another procedure is preferred, in which event it is well to state precisely the method followed. The student will find, of course, as he progresses with analytical

studies that sometimes a straight line of better fit can be obtained by varying from this general rule, so that any such generalization as has been formulated may safely be modified when the fundamental principles of regression are well understood. It has been the experience of the author that in many instances an exceptionally high coefficient of regression can be obtained by using the coefficient of correlation as calculated by the method of percentage change of first differences and the standard deviations as calculated from the squares of deviations of actual values from the mean of actual values. Of course, an equally high, or even higher, degree of regression in other instances might be obtained from other combinations of correlation and standard deviation. The problem of fitting a straight line of relationship, then, is one that will be decided upon its own individual merits, and the purpose for which the regression chart is being constructed.

In common parlance we hear the expressions of regression of  $y$  on  $x$  and regression of  $x$  on  $y$ . Perhaps it may be well to inquire into the exact meaning implied by these expressions. A regression coefficient of  $-.82$  obtained by the equation  $b = r \frac{\sigma_y}{\sigma_x}$ , in which  $r$  is the Pearsonian coefficient of correlation, indicates that for each change of one unit in the  $x$ , or independent, variable from the average change there is a corresponding opposite change in the  $y$ , or dependent, variable of  $.82$  of one unit. Therefore, if in calculating the coefficient of correlation between cotton production and cotton prices the  $x$  variable, production, is expressed in millions of bales, and the  $y$  variable, price, is expressed in cents per pound, a regression coefficient of  $-.82$  based on the coefficient of correlation thus calculated would indicate that for each change in production of one

million bales from the average production, the most probable change in price from the average price is .82 of one cent in the opposite direction. From this we see that price ( $y$ ) changes with production ( $x$ ). With each change of one unit in production there was a change in price in the opposite direction of only .82 of one unit. Thus it is seen that the change in price, the dependent variable, was of less magnitude than the change in production, which is the independent value. The nature of the relationship between corresponding pairs of items in the two series is such that, as indicated, an increase or decrease in production is generally accompanied or followed by a change in price in the opposite direction, so that the price with each change, whether the change be plus or minus, always regresses, or tends to regress, toward the average price. We have, therefore, the regression of  $y$  on  $x$ . On the other hand, however, if we were calculating the regression coefficient

from  $b = r \frac{\sigma x}{\sigma y}$  and obtained a value of 1.20 the inference

might at first be drawn that a change of one cent per pound in price from the average change is associated with a change in production in the opposite direction of 1.20 millions of bales from the average year-to-year change. This, then, would be the regression of  $x$  on  $y$ . The meanings of the two, however, must not be confused, and

neither must it be concluded that the equations  $b = r \frac{\sigma y}{\sigma x}$  and  $b = r \frac{\sigma x}{\sigma y}$  are interchangeable. If we are calculating

the coefficient of correlation between cotton prices for December and cotton production for the period, say, from 1920 to 1928, it is obvious that we would use the data for those particular years. We would then have an in-

dication of the effect of production on price because the crop of the current year would have exerted most of its causal effect or influence upon the December price.

Clearly, then,  $b = r \frac{\sigma_y}{\sigma_x}$  gives the slope of the line of high-

est linear relationship between the two series. But, suppose it is desired to calculate the slope of the line of highest linear relationship between two series with production ( $x$ ) being considered as dependent upon price. In this instance it would be manifestly incorrect to divide the standard deviation of production for the years 1920 to 1928 by the standard deviation of December price for those same years. This is because the production of cotton in any one year is not affected by the December price of that year, except that harvestings may not be complete if prices are low. It is the production of the following year that will likely be affected by price, the assumption being that yield per acre will remain about the same, because of the tendency for acreage planted in the spring to vary in the same direction as the change in preceding prices. Therefore, in calculating the regression of production or acreage on price the logical procedure in the preceding illustration would be to use December prices for the years 1920 to 1928 and production or acreage data for the years 1921 to 1929. In actual practice, of course, the usual procedure is to consider the dependent variable as the  $y$  variable, so that the regression coefficient is generally expressed as the regression of  $y$  on  $x$ . True regression of  $x$  on  $y$  is exemplified when the dependent variable is represented on the  $x$  axis and the independent variable on the  $y$  axis. Accordingly, we may plot in a scatter diagram, for example, the relationship between price of cotton in December and the acreage of cotton planted in the

following spring and still represent the acreage on the x axis.

In Chart 22, Chapter IV, is shown the straight line relationship between precipitation and yield of peanuts as represented in Tables XXXVI and XXXVII. This line, TT, is so drawn that the sum of the squares of the deviations from it is less than the sum of the squares of the deviations from any other line similarly drawn. It may be known, therefore, as the line of best fit, or the line of least squares. The significance of this line is that it shows the closest relationship between precipitation and yield of peanuts in Tables XXXVI and XXXVII that can be represented by a straight line.

The slope of the straight line in Chart 22 is calculated by means of the following regression equation:

$b = r \frac{\sigma y}{\sigma x}$ , in which  $b$  is the coefficient of regression,  $r$  the

coefficient of correlation between precipitation and yield,  $\sigma y$  the standard deviation of yield, and  $\sigma x$  the standard deviation of precipitation. By the Pearsonian method of correlation the value of  $r$  was found to be  $-.75$ . The standard deviation of yield is 143.10, and the standard deviation of precipitation is 2.10, while the coefficient of regression, or the slope of the line of highest average relationship, is  $-51.10$ . As shown in Tables XXXVI and XXXVII, the respective means are 4.58 and 849. The intersection of these means is the ordinate of the straight line at that point. The other ordinates of the line are located by means of the equation  $y = a + bx$ , in which  $a$  is the value on the y axis where the two means intersect,  $b$  the coefficient of regression, and  $x$  the distance between values on the x axis. The regression coefficient,  $-51.10$ , indicates that for each change in precipitation of one inch



above or below the average the most probable change in yield of peanuts per acre is a variation in the opposite direction to the extent of 51.10 pounds.

Simple regression may be curvilinear as well as rectilinear, since, as has already been explained, the relationship between a dependent variable and a single series of independent variables may be such that it cannot be represented by a straight line. In calculating the coefficient of simple curvilinear regression the correlation index or the correlation ratio would be used instead of the coefficient of correlation.

### MULTIPLE REGRESSION

The method of calculating a numerical expression of combined effect of two or more independent variables upon a dependent has already become familiar from the study of Chapter XII. It remains now to inquire into the method of determining the most probable value of the dependent variable when two or more independents are considered. At the outset it may be well to bear in mind that just as the formula  $y = a + bx$  for a straight line gives ordinates that are identical with those of the simple regression line, so the multiple regression equation leads to the solving for values of the dependent variable that are identical with the values most likely to be associated with given values of the independents. It is in this sense that the term "multiple regression" is applied, since the dependent variable is, in effect, the culminating or resulting product of the multiple effect of two or more causal factors. Thus it is that the multiple regression equation is the same as the estimating equation based on multiple relationship.

For illustrative purposes the application of the multiple



regression equation will be based on the data in the following table.

TABLE LXXX. WEATHER CONDITIONS AT COLLEGE PARK AND WHEAT YIELDS IN MARYLAND, 1909-1926

| 1                  | 2                                    | 3                 | 4                 | 5                                 | 6                                                 |
|--------------------|--------------------------------------|-------------------|-------------------|-----------------------------------|---------------------------------------------------|
| YEAR               | WEATHER CONDITIONS AT COLLEGE PARK * |                   |                   | WHEAT YIELDS PER ACRE IN MARYLAND |                                                   |
|                    | Snow Cover in Late Winter †          | Sunshine in May † | Rainfall in May † | Actual Yield per Acre ‡           | Ordinates of Straight Line Trend of Least Squares |
|                    | <i>Inches</i>                        | <i>Hours</i>      | <i>Inches</i>     | <i>Bushels</i>                    | <i>Bushels</i>                                    |
| 1909               | 6                                    | 279               | 4.86              | 14.5                              | 14.8                                              |
| 1910               | 6                                    | 283               | 4.97              | 17.4                              | 15.1                                              |
| 1911               | 1                                    | 355               | .96               | 15.5                              | 15.3                                              |
| 1912               | 18                                   | 284               | 3.44              | 15.0                              | 15.5                                              |
| 1913               | 20                                   | 219               | 4.87              | 13.3                              | 15.7                                              |
| 1914               | 22                                   | 343               | 1.04              | 21.5                              | 16.0                                              |
| 1915               | 4                                    | 240               | 6.10              | 16.1                              | 16.2                                              |
| 1916               | 7                                    | 276               | 4.46              | 16.0                              | 16.4                                              |
| 1917               | 4                                    | 272               | 3.44              | 17.0                              | 16.7                                              |
| 1918               | 13                                   | 305               | 3.99              | 15.5                              | 16.9                                              |
| 1919               | 0                                    | 244               | 5.07              | 13.5                              | 17.1                                              |
| 1920               | 12                                   | 279               | 3.62              | 17.0                              | 17.3                                              |
| 1921               | 4                                    | 226               | 4.42              | 14.0                              | 17.6                                              |
| 1922               | 16                                   | 289               | 3.89              | 16.5                              | 17.8                                              |
| 1923               | 4                                    | 329               | 2.05              | 19.2                              | 18.0                                              |
| 1924               | 10                                   | 244               | 7.61              | 15.8                              | 18.2                                              |
| 1925               | 7                                    | 311               | 1.86              | 21.0                              | 18.5                                              |
| 1926               | 15                                   | 302               | 3.34              | 23.0                              | 18.7                                              |
| Mean               | 8.8                                  | 282.2             | 3.89              | 16.8                              | —                                                 |
| Standard deviation | 6.07                                 | 37.43             | 1.64              | 2.68                              | —                                                 |

\* College Park is situated near Washington, D. C., in Prince George's County, Maryland, on the Washington-Baltimore boulevard. It is there that is located Maryland's agricultural experiment station.

† Compiled from records of the Maryland Agricultural Experiment Station.

‡ Yearbooks, United States Department of Agriculture, 1928, page 675, Table 6, and 1924, page 563, Table 6.

In columns 2, 3, and 4 of the table are shown the statistics of specified prevailing weather conditions at College Park, Maryland, and the statistics of yields of wheat in

Maryland that followed the occurrence of those particular conditions are shown in column 5. Our problem is to determine what yield would be expected to be associated with given weather conditions. For this purpose the following multiple regression equation will be found very satisfactory:

$$y = pax \frac{a - \hat{a}}{SDa} + pbx \frac{b - \hat{b}}{SDb} + pcx \frac{c - \hat{c}}{SDc}.$$

The meanings of the symbolic expressions are as follows:

$y$  = the value of the dependent variable most likely to be associated with given values of the independents

$pax$  = the path coefficient between snow cover in late winter and yield of wheat

$a - \hat{a}$  = the difference between snow cover in any one year and the average of snow cover for all the years

$SDa$  = the standard deviation of snow cover

$pbx$  = the path coefficient between sunshine in May and yield of wheat

$b - \hat{b}$  = the difference between the hours of sunshine in May for any one year and the average of hours of sunshine in May for all the years

$SDb$  = the standard deviation of May sunshine

$pcx$  = the path coefficient between rainfall in May and yield of wheat

$c - \hat{c}$  = the difference between rainfall in May for any one year and the average of rainfall in May for all the years

$SDc$  = the standard deviation of rainfall.

In the calculation of the coefficient of multiple correlation between the three independent variables and yield,

Chapter XII, it was found that the path coefficient between snow cover in late winter and yield is .318,<sup>1</sup> while it is .942 between May sunshine and yield, and .440 between rainfall in May and yield. Table LXXX shows the means and the standard deviations of the dependent variable and the three independents, so that the multiple regression equation is readily applied.

Let us calculate from the multiple regression equation the yield of wheat that would be expected to be associated with snow cover in late winter, sunshine in May, and rainfall in May for the year 1914. Proceeding with this calculation we have:

$$y = pax \frac{a - \hat{a}}{SDa} + pbx \frac{b - \hat{b}}{SDb} + pcx \frac{c - \hat{c}}{SDc},$$

which by substitution equals

$$\begin{aligned} y &= .318 \frac{22 - 8.8}{6.07} + .942 \frac{343 - 282.2}{37.43} + .440 \frac{1.04 - 3.89}{1.64} \\ &= .692 + (1.530) + (-.765) \\ &= .692 + 1.530 - .765 = 1.457. \end{aligned}$$

The value of .692 is obtained by subtracting 8.8, the mean of snow cover, from 22, the snow cover for late winter of 1914, dividing the remainder by 6.07, the standard deviation of snow cover, and then multiplying by .318, the path coefficient between snow cover and yield. The next value, 1.530, is obtained by subtracting 282.2, the mean of sunshine, from 343, the hours of sunshine for May, 1914, dividing the remainder by 37.43, the standard deviation of sunshine, and then multiplying by .942, the path coefficient between sunshine and yield.

<sup>1</sup> This is the simple coefficient of correlation between snow cover in late winter and yield of wheat. No interrelationship is calculated between snow cover in late winter and rainfall in May. See the discussion of multiple correlation in Chapter XII, pages 225 to 228, that is based on the data in Table LXXX.

The negative value,  $-.765$ , is obtained by subtracting 3.89, the mean of rainfall in May, from 1.04, the rainfall in May of 1914, dividing the remainder by 1.64, the standard deviation of rainfall, and then multiplying by .440, the path coefficient between rainfall and yield. Since the mean of rainfall is greater than the rainfall for May, 1914, it necessarily follows that when the mean is subtracted a minus quantity remains. It is for this reason that the value of  $-.765$  carries the minus sign.

The significance of the values obtained by solving the multiple regression equation after proper substitution is as follows:

The value of .692 means that the most probable yield of wheat in 1914 is that part of a standard deviation above the average yield due to snow cover in late winter.

The value of 1.530 means that the most probable yield of wheat in 1914 is that many standard deviations greater than the average yield due to sunshine in May.

The minus value of .765 means that the most probable yield of wheat in 1914 is that part of a standard deviation less than the average yield due to rainfall in May.

In other words, the indications are that snow cover and sunshine acted favorably on yield in 1914, whereas rainfall acted unfavorably. The net effect of these three factors is obtained by summing .692, 1.530, and  $-.765$ , which gives a total of 1.457. This means that, on the basis of the influence of these three weather factors, the most probable yield of wheat for 1914 is 1.457 standard deviations greater than the average yield for the period from 1909 to 1926. We have already seen in Table LXXX

that the mean of yield is 16.8 bushels, and that the standard deviation is 2.68. Multiplying 2.68 by 1.457 gives 3.9. Adding 3.9 to 16.8 we obtain a value of 20.7, the yield most likely to have been expected in 1914. The actual yield for 1914 will be seen to have been 21.5 bushels, which is .8 of a bushel greater than the most probable, or estimated, yield. The other most probable yields for all the other years are calculated in the same way by applying the multiple regression equation and making the proper substitutions.

Attention may be called to the fact that if in the multiple regression equation all the deviations are measured from trend lines instead of from arithmetic means, the value obtained by solving indicates that the most probable yield of wheat for any one year is to be determined from the ordinate of trend in yield for that year, rather than from the mean of yield for the period.

## CHAPTER XIV

### ESTIMATES AND ERROR

The fundamental principle involved in the calculation of numerical expressions of cause and effect relationships is that with a determination of the extent to which dependent and independent variables are associated, or have been associated in the past, a conception of probable future relationships may be formed. When we have measured the degree of association that exists, or that has existed, between a dependent variable and an influencing factor we are in a position to estimate or predict the most probable magnitude of the associated dependent variable. This is the principle underlying all problems of forecasting,<sup>1</sup> for it is only on the basis of past occurrences that we can arrive at any conclusion as to what is most likely to happen in the future.

The discussion of regression in the preceding chapter has already illustrated in a rather clear way the method of estimating most probable values of dependent variables when the magnitudes of the independents are known. The simple regression line, for example, enables us to determine the ordinate of the straight line trend of relationship between variables in two series when the size of the independent variable is known. Thus it is that the most probable price of cotton for any month can be projected in advance as soon as the production of the cur-

<sup>1</sup> In this discussion the terms "forecasting" and "prediction" have reference to future probabilities, while "estimates" and "estimating" relate to past occurrences.



rent year is fairly well known. In like manner, the multiple regression equation provides a means of predicting the most probable value of the dependent variable when the magnitudes of independents are known. It is in this way that a forecast may be made of probable wheat yields, for instance, when the effects of preceding weather conditions have had opportunity to manifest themselves.

Prediction from regression equations is merely a continuation of the making of estimates already illustrated in Chapter XIII, so that further discussion is not deemed necessary. The principle is briefly reviewed in this chapter in order that the student may realize more fully the added significance attached to regression lines. It is especially to be remembered that, theoretically at least, any regression equation that can be used for estimating values that would have been expected to be associated with certain causal factors can be used with almost equal propriety for actually predicting in the future.

### THE ESTIMATING EQUATION

Perhaps a more satisfactory method of making predictions involves taking into account the fact that the dependent variable under actual conditions can seldom be expected to coincide with the ordinate of a calculated trend of relationship, or the ordinate of the regression line, just as with the independent variable. Let us take for illustrative purposes a dependent series as influenced by a single independent series. By the ordinary regression equation the most probable dependent value corresponding to a given independent value is calculated as the ordinate of straight line trend of highest relationship, whereas the actual value of the dependent may deviate considerably from the ordinate of the calculated

trend of relationship. Our problem is to improve this method of making estimates. Recalling that all estimating and predictive equations are based upon cause and effect relationships that have existed in the past, we will proceed with a practical problem to illustrate a commonly accepted procedure. By the method of correlation of percentage change of first differences between the average of spot quotations for middling upland cotton for the month of December, 1906 to 1923, and the acreage of cotton harvested in the United States in the years from 1907 to 1924, a coefficient of correlation of .80 was calculated.<sup>1</sup> This coefficient is used in calculating the coefficient of regres-

sion by the formula  $b = r \frac{\sigma y}{\sigma x}$ , which becomes a com-

ponent factor in the estimating or predictive equation. This equation may be expressed as  $y = Ay - bAx + bx$ . Applying it to the acreage of cotton harvested in any year as related to preceding December cotton prices the meanings of the symbolic expressions are as follows:

$y$  = the percentage change in acreage

$Ay$  = the average of percentage change in acreage

$b$  = the coefficient of regression of acreage on price

$x$  = the percentage change in price

$Ax$  = the average of percentage change in price.

The calculation of the estimated values of the dependent variable, acreage, is now a simple matter. It is done by merely subtracting from the average of percentage changes

<sup>1</sup> In the calculation of this coefficient of correlation the averages of spot quotations for middling upland cotton at New York and New Orleans as reported in *Yearbook, United States Department of Agriculture, 1923*, page 809, Table 307, were deflated with index numbers of farm products prices published by the Bureau of Labor Statistics in *Index Numbers of Wholesale Prices, by Years and Months, 1890 to August, 1926*, pages 4 to 8. These deflated prices were then correlated with the acreages of cotton harvested in the United States as reported in *Yearbook, United States Department of Agriculture, 1928*, page 837, Table 250.

in acreage ( $Ay$ ) the product of the coefficient of regression and the average of percentage change in price and the product of the coefficient of regression and the percentage change in price, which is but the application of the formula  $y = Ay - bAx + bx$ . Table LXXXI shows the actual percentage changes in cotton acreage harvested for a period of years and the percentage changes as estimated by means of the above formula.

TABLE LXXXI. ACTUAL AND ESTIMATED PERCENTAGE CHANGES IN COTTON ACREAGE HARVESTED IN THE UNITED STATES, 1908-1924

| YEAR    | ACTUAL PERCENTAGE<br>CHANGE IN ACREAGE OF<br>COTTON HARVESTED * | ESTIMATED PERCENTAGE<br>CHANGE IN COTTON<br>ACREAGE HARVESTED † |
|---------|-----------------------------------------------------------------|-----------------------------------------------------------------|
|         | <i>Per Cent.</i>                                                | <i>Per Cent.</i>                                                |
| 1908    | 9.39                                                            | 6.85                                                            |
| 1909    | - 4.64                                                          | - 4.33                                                          |
| 1910    | 4.74                                                            | 11.69                                                           |
| 1911    | 11.24                                                           | - .78                                                           |
| 1912    | - 4.89                                                          | - 6.13                                                          |
| 1913    | 8.18                                                            | 6.85                                                            |
| 1914    | - .69                                                           | 1.21                                                            |
| 1915    | -14.72                                                          | - 9.13                                                          |
| 1916    | 11.37                                                           | 14.52                                                           |
| 1917    | - 3.27                                                          | 6.89                                                            |
| 1918    | 6.40                                                            | 1.99                                                            |
| 1919    | - 6.78                                                          | - 2.22                                                          |
| 1920    | 6.89                                                            | 6.35                                                            |
| 1921    | -14.96                                                          | -12.33                                                          |
| 1922    | 8.28                                                            | 3.08                                                            |
| 1923    | 12.37                                                           | 8.17                                                            |
| 1924    | 11.41                                                           | 7.53                                                            |
| Average | 2.37                                                            | 2.36                                                            |

\* Calculated from acreages reported in *Yearbook, United States Department of Agriculture, 1928*, page 837, Table 250.

† Estimated from the equation  $y = Ay - bAx + bx$ , in which  $b$  is calculated from the coefficient of correlation of .80 between the average of deflated prices for middling upland cotton at New York and New Orleans for December, 1906 to 1923, and the acreage of cotton harvested in the United States during the crop years 1907 to 1924.

As will be observed in the table, the average of percentage changes in acreage as estimated from the numerical expression of relationship between the average of

December deflated spot quotations for middling upland cotton at New York and New Orleans and the acreage of cotton harvested in the subsequent crop years differs only one one-hundredth of one per cent. from the average of actual percentage changes. This may be taken to indicate that when the December price is known a fairly accurate estimate may be made of the percentage change in acreage of cotton that is likely to be harvested in the next crop season.

The estimated percentage change in acreage being calculated, the estimate of actual acreage is made by multiplying the acreage of any year immediately preceding

TABLE LXXXII. ACTUAL AND ESTIMATED ACREAGES OF COTTON  
HARVESTED IN THE UNITED STATES, 1908-1924

| 1     | 2                             | 3                                   |
|-------|-------------------------------|-------------------------------------|
| YEAR  | ACTUAL ACREAGE<br>HARVESTED * | ESTIMATES OF ACREAGE<br>HARVESTED † |
|       | <i>1,000 Acres</i>            | <i>1,000 Acres</i>                  |
| 1908  | 32,444                        | 31,690                              |
| 1909  | 30,938                        | 31,039                              |
| 1910  | 32,403                        | 34,555                              |
| 1911  | 36,045                        | 32,150                              |
| 1912  | 34,283                        | 33,835                              |
| 1913  | 37,089                        | 36,631                              |
| 1914  | 36,832                        | 37,538                              |
| 1915  | 31,412                        | 33,469                              |
| 1916  | 34,985                        | 35,973                              |
| 1917  | 33,841                        | 37,395                              |
| 1918  | 36,008                        | 34,514                              |
| 1919  | 33,566                        | 35,209                              |
| 1920  | 35,878                        | 35,697                              |
| 1921  | 30,509                        | 31,454                              |
| 1922  | 33,036                        | 31,449                              |
| 1923  | 37,123                        | 35,735                              |
| 1924  | 41,360                        | 39,918                              |
| Total | 587,752                       | 588,251                             |
| Mean  | 34,544                        | 34,603                              |

\* *Yearbook, United States Department of Agriculture, 1928*, page 837, Table 250.

† Based on estimated percentage changes in Table LXXXI, page 265.

the year for which the acreage is to be estimated by the percentage change and then adding or subtracting the product, as the case may be. For example, the estimated acreage in 1920 was found to be 6.35 per cent. greater than the acreage of 1919. Table LXXXII shows the acreage in 1919 to have been 33,566,000 acres. By taking 6.35 per cent. of 33,566,000 and adding it to 33,566,000, the estimate of the 1920 acreage is obtained, which is 35,697,000, as compared with an actual acreage of 35,878,000. If the percentage change had been a minus quantity a subtraction would have been made instead of an addition. Table LXXXII shows statistics of the actual acreages of cotton harvested from 1908 to 1924 and the acreage figures as estimated from the percentage changes in the third column of Table LXXXI.

The average of the difference between the total acreage harvested from 1908 to 1924 and the acreage as estimated is approximately 29,000 acres. On the basis of the weighted average yield of lint cotton per acre in the United States for that period these 29,000 acres are equivalent to about 10,000 bales of cotton lint. In estimating, a slight variation like this is negligible, since a difference as great as 2,000,000 acres would have been equivalent to only 700,000 bales of lint. The extent to which the estimated acreage approximates the actual acreage will become even more significant when it is recalled that estimates are made on the basis of December price one year in advance of the time when practically all the cotton is harvested. From this it is indicated that a fairly close prediction may be made in late December or early January as to the acreage of cotton that is likely to be planted and harvested during the subsequent crop year. With this predicted acreage at hand, an estimate may then be made

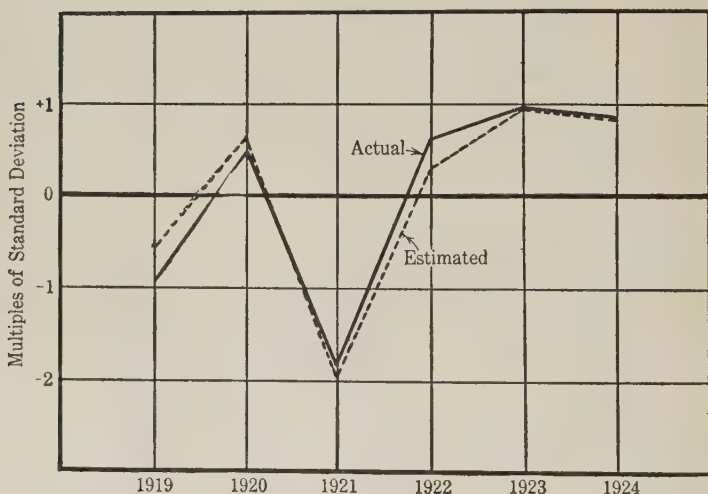


of probable production, and finally of prices that are likely to be received during the months following harvest.

It is probably well to call attention to one significant feature of the equation used in estimating the values in Table LXXXI. The coefficient of correlation between price and acreage, we recall, was calculated by the method of percentage change of first differences. By thus calculating the numerical expression of relationship, the deviations of

CHART 40. ACTUAL AND ESTIMATED PERCENTAGE CHANGES IN COTTON ACREAGE HARVESTED IN THE UNITED STATES, 1919-1924

(Based on data in Table LXXXI)



percentage changes of first differences from their mean were reduced to terms of multiples of standard deviation. In constructing the estimating equation the coefficient of regression converts the multiples of standard deviation back into actual values. This is why the equation  $y = Ay - bAx + bx$  gives a value in the same terms as that of the data correlated.



The preceding chart is presented to furnish a clearer conception of the degree of approach to absolute accuracy that may sometimes accompany the making of estimates, and to further indicate the practical application that might possibly be made of predictive equations.

#### RATIO METHOD OF ESTIMATING <sup>1</sup>

To obviate the necessity of constructing forecasting equations involving the calculation of correlation and regression coefficients, it is sometimes desirable that some other methods be devised that will satisfactorily serve the same purpose. One of the best of these is what may be termed the "ratio method" of prediction. In arriving at the most probable value of the dependent variable by this method, it is only necessary to express the independent variables in terms of ratios, obtain their product, and then multiply by the dependent value immediately preceding the one to be ascertained. To illustrate this method of estimating let us refer to Table LXXXIII. The problem here is to estimate the production of cotton in the United States on the basis of past yields and current acreage. In column 2 are the average yields per acre for the years from 1899 to 1926. Column 3 shows the weighted average yields for nine-year periods, while in column 4 we have the yield ratio, or the weighted average yields in column 3 expressed as ratios to the yields in column 2. The acreage ratios are shown in column 5. These ratios are merely the acreages of the specified years in column 1 expressed as ratios to the acreages of immediately preceding years. The statistics of cotton production for the years preceding those in column 1 are given in column 6. All that is necessary to

<sup>1</sup> Acknowledgments are made to Bradford B. Smith of the Cleveland Trust Company, Cleveland, Ohio, formerly of the United States Department of Agriculture.

TABLE LXXXIII. COTTON PRODUCTION ESTIMATES IN

| 1    | 2                                                                     | 3                                                                                                                                   | 4                                                     | 5                                                                                               |
|------|-----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| YEAR | YIELD OF COTTON<br>PER ACRE FOR<br>YEAR PRECEDING<br>SPECIFIED YEAR * | WEIGHTED AVER-<br>AGE YIELD OF<br>COTTON PER ACRE<br>FOR NINE-YEAR<br>PERIOD ENDING<br>WITH YEAR PRE-<br>CEDING SPECIFIED<br>YEAR † | YIELD RATIO<br>(Column 3<br>Divided by<br>Column 2) ‡ | ACREAGE RATIO<br>(Acreage of<br>Specified Year<br>Divided by<br>Acreage of<br>Preceding Year) ‡ |
|      | <i>Pounds</i>                                                         | <i>Pounds</i>                                                                                                                       |                                                       |                                                                                                 |
| 1900 | 183.8                                                                 | 195.5                                                                                                                               | 1.064                                                 | 1.025                                                                                           |
| 1901 | 194.4                                                                 | 192.2                                                                                                                               | .989                                                  | 1.074                                                                                           |
| 1902 | 170.0                                                                 | 188.8                                                                                                                               | 1.111                                                 | 1.015                                                                                           |
| 1903 | 187.3                                                                 | 189.3                                                                                                                               | 1.011                                                 | .995                                                                                            |
| 1904 | 174.3                                                                 | 186.9                                                                                                                               | 1.072                                                 | 1.154                                                                                           |
| 1905 | 205.9                                                                 | 191.2                                                                                                                               | .929                                                  | .868                                                                                            |
| 1906 | 186.6                                                                 | 192.2                                                                                                                               | 1.030                                                 | 1.157                                                                                           |
| 1907 | 202.5                                                                 | 191.2                                                                                                                               | .944                                                  | .945                                                                                            |
| 1908 | 179.1                                                                 | 187.4                                                                                                                               | 1.046                                                 | 1.094                                                                                           |
| 1909 | 194.9                                                                 | 188.8                                                                                                                               | .969                                                  | .954                                                                                            |
| 1910 | 154.3                                                                 | 184.0                                                                                                                               | 1.192                                                 | 1.047                                                                                           |
| 1911 | 170.7                                                                 | 184.0                                                                                                                               | 1.078                                                 | 1.112                                                                                           |
| 1912 | 207.7                                                                 | 186.9                                                                                                                               | .900                                                  | .951                                                                                            |
| 1913 | 190.9                                                                 | 188.8                                                                                                                               | .989                                                  | 1.082                                                                                           |
| 1914 | 182.0                                                                 | 185.9                                                                                                                               | 1.021                                                 | .993                                                                                            |
| 1915 | 209.2                                                                 | 188.8                                                                                                                               | .902                                                  | .853                                                                                            |
| 1916 | 170.3                                                                 | 185.5                                                                                                                               | 1.089                                                 | 1.114                                                                                           |
| 1917 | 156.6                                                                 | 182.6                                                                                                                               | 1.166                                                 | .967                                                                                            |
| 1918 | 159.7                                                                 | 178.8                                                                                                                               | 1.120                                                 | 1.064                                                                                           |
| 1919 | 159.6                                                                 | 179.3                                                                                                                               | 1.123                                                 | .932                                                                                            |
| 1920 | 161.5                                                                 | 178.3                                                                                                                               | 1.104                                                 | 1.069                                                                                           |
| 1921 | 178.4                                                                 | 174.9                                                                                                                               | .980                                                  | .850                                                                                            |
| 1922 | 124.5                                                                 | 168.3                                                                                                                               | 1.352                                                 | 1.083                                                                                           |
| 1923 | 141.2                                                                 | 163.5                                                                                                                               | 1.158                                                 | 1.124                                                                                           |
| 1924 | 130.6                                                                 | 153.9                                                                                                                               | 1.178                                                 | 1.114                                                                                           |
| 1925 | 157.4                                                                 | 153.0                                                                                                                               | .972                                                  | 1.113                                                                                           |
| 1926 | 167.2                                                                 | 154.4                                                                                                                               | .923                                                  | 1.022                                                                                           |
| 1927 | 182.6                                                                 | 157.7                                                                                                                               | .864                                                  | .852                                                                                            |

\* *Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.*

† Based on acreage harvested and total production as reported in *Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.* Average yield in fractional part of a bale converted to pound equivalents on the basis of 478 pounds of lint cotton per bale.

‡ Based on acreage harvested in the United States as reported in *Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.*

estimate the probable production for any year is to obtain the product of the yield ratio in column 4 and the acreage ratio in column 5 and then multiply by the production of cotton for the year immediately preceding the specified year. For instance, the yield and acreage ratios opposite

## THE UNITED STATES BY THE RATIO METHOD, 1900-1927

| 6                                                                 | 7                                               | 8                                                                                                                                                    | 9                                                                                                                               |
|-------------------------------------------------------------------|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| PRODUCTION OF<br>COTTON FOR YEAR<br>PRECEDING<br>SPECIFIED YEAR * | PRODUCTION OF<br>COTTON FOR<br>SPECIFIED YEAR * | ESTIMATED PRODUCTION OF COTTON<br>FOR SPECIFIED<br>YEAR (Production<br>of Preceding Year<br>Times the Product<br>of the Yield and<br>Acreage Ratios) | EXTENT TO WHICH<br>THE ESTIMATED<br>PRODUCTION DIFFERS FROM THE<br>ACTUAL PRODUCTION<br>(Difference between<br>Columns 7 and 8) |
| <i>1,000 Bales</i>                                                | <i>1,000 Bales</i>                              | <i>1,000 Bales</i>                                                                                                                                   | <i>1,000 Bales</i>                                                                                                              |
| 9,345                                                             | 10,123                                          | 10,192                                                                                                                                               | 69                                                                                                                              |
| 10,123                                                            | 9,510                                           | 10,753                                                                                                                                               | 1,243                                                                                                                           |
| 9,510                                                             | 10,631                                          | 10,724                                                                                                                                               | 93                                                                                                                              |
| 10,631                                                            | 9,851                                           | 10,694                                                                                                                                               | 843                                                                                                                             |
| 9,851                                                             | 13,438                                          | 12,187                                                                                                                                               | -1,251                                                                                                                          |
| 13,438                                                            | 10,575                                          | 10,836                                                                                                                                               | 261                                                                                                                             |
| 10,575                                                            | 13,274                                          | 12,602                                                                                                                                               | - 672                                                                                                                           |
| 13,274                                                            | 11,107                                          | 11,841                                                                                                                                               | 734                                                                                                                             |
| 11,107                                                            | 13,242                                          | 12,710                                                                                                                                               | - 532                                                                                                                           |
| 13,242                                                            | 10,005                                          | 12,241                                                                                                                                               | 2,236                                                                                                                           |
| 10,005                                                            | 11,609                                          | 12,486                                                                                                                                               | 877                                                                                                                             |
| 11,609                                                            | 15,693                                          | 13,916                                                                                                                                               | -1,777                                                                                                                          |
| 15,693                                                            | 13,703                                          | 13,432                                                                                                                                               | - 271                                                                                                                           |
| 13,703                                                            | 14,156                                          | 14,664                                                                                                                                               | 508                                                                                                                             |
| 14,156                                                            | 16,135                                          | 14,352                                                                                                                                               | -1,783                                                                                                                          |
| 16,135                                                            | 11,192                                          | 12,414                                                                                                                                               | 1,222                                                                                                                           |
| 11,192                                                            | 11,450                                          | 13,578                                                                                                                                               | 2,128                                                                                                                           |
| 11,450                                                            | 11,302                                          | 12,910                                                                                                                                               | 1,608                                                                                                                           |
| 11,302                                                            | 12,041                                          | 13,468                                                                                                                                               | 1,427                                                                                                                           |
| 12,041                                                            | 11,421                                          | 12,603                                                                                                                                               | 1,182                                                                                                                           |
| 11,421                                                            | 13,440                                          | 13,479                                                                                                                                               | 39                                                                                                                              |
| 13,440                                                            | 7,954                                           | 11,196                                                                                                                                               | 3,242                                                                                                                           |
| 7,954                                                             | 9,755                                           | 11,646                                                                                                                                               | 1,891                                                                                                                           |
| 9,755                                                             | 10,140                                          | 12,697                                                                                                                                               | 2,557                                                                                                                           |
| 10,140                                                            | 13,628                                          | 13,307                                                                                                                                               | - 321                                                                                                                           |
| 13,628                                                            | 16,104                                          | 14,743                                                                                                                                               | -1,361                                                                                                                          |
| 16,104                                                            | 17,977                                          | 15,191                                                                                                                                               | -2,786                                                                                                                          |
| 17,977                                                            | 12,955                                          | 13,233                                                                                                                                               | 278                                                                                                                             |

\* Yearbook, United States Department of Agriculture, 1928, page 837, Table 250.

the year 1900 are 1.064 and 1.025, respectively. The weighted average yield for the nine-year period ending in 1899 is 1.064 times as great as the yield for 1899, while the acreage of 1900 is 1.025 times that of 1899. Multiplying 1.064 by 1.025 gives a product of 1.0906, which, when multiplied by 9,345,000, the production for 1899, gives 10,192,000, the estimated production for 1900.

For the entire twenty-eight year period represented in the table the average of estimated production was 12,-646,000, whereas the average of actual production was 12,-229,000. The great defect in this method of estimating, as in all other methods, is that individual estimates may sometimes deviate greatly from actual conditions. However, for much of the ordinary prediction work in which two, or even more, independent variables are used it will probably be found that the ratio method has as great application as an equation formulated on the basis of calculated interrelationships. It is well, though, for the student to realize that when there are numerous causal factors exerting their influences the reliability of any predictive equation becomes problematical. The value of multiple regression in forecasting is probably sometimes over-emphasized, particularly in regard to strictly economic data.

#### PAR FORECASTING <sup>1</sup>

The necessary limitations that must be placed upon the rigid mathematical predictive equation when there is a wide diversity of causal factors, with consequent variations in effect upon response groups, is probably better known to the agricultural specialist than to any other worker. It is he, more than any other, who must forecast the probable effect of a multitude of factors, giving varying weights to these factors from time to time, and forming his best judgment as to final outcome. Experience has shown to considerable extent that when a forecast is to be made during the growing season of the probable production of a particular crop the safest and most logical plan is to base this forecast on the condition of the crop at

<sup>1</sup> Information on the making of par forecasts may be obtained without cost from the Crop Reporting Board, United States Department of Agriculture, Washington, D. C.

the time, and the condition and subsequent production of former years. Making of production forecasts on this basis at intervals during the growing season is likely to approach a much higher degree of accuracy than those made from rigid equations. The condition par takes into account the effects of improvements in cultural practices, influences due to late or early plantings, storm damage, ravages of diseases and insect pests, and countless other factors which do not readily lend themselves to precise numerical measurement.

To illustrate the approach to accuracy that may be attained by the par method of forecasting, let us refer to the work of the United States Department of Agriculture on cotton production. In Table LXXXIV are shown the forecasts of cotton production in the United States for

TABLE LXXXIV. PRODUCTION OF COTTON IN THE UNITED STATES AND PAR FORECASTS FOR SPECIFIED MONTHS, 1915-1927

| YEAR    | PRODUCTION<br>OF COTTON * | PAR FORECASTS OF PRODUCTION † |                    |                    |                    |
|---------|---------------------------|-------------------------------|--------------------|--------------------|--------------------|
|         |                           | July                          | August             | September          | October            |
|         | <i>1,000 Bales</i>        | <i>1,000 Bales</i>            | <i>1,000 Bales</i> | <i>1,000 Bales</i> | <i>1,000 Bales</i> |
| 1915    | 11,192                    | 12,381                        | 11,876             | 11,697             | 10,950             |
| 1916    | 11,450                    | 14,266                        | 12,916             | 11,800             | 11,637             |
| 1917    | 11,302                    | 11,633                        | 11,949             | 12,499             | 12,047             |
| 1918    | 12,041                    | 15,327                        | 13,619             | 11,137             | 11,818             |
| 1919    | 11,421                    | 10,986                        | 11,016             | 11,230             | 10,696             |
| 1920    | 13,440                    | 11,450                        | 12,519             | 12,783             | 12,123             |
| 1921    | 7,954                     | 8,433                         | 8,203              | 7,037              | 6,537              |
| 1922    | 9,755                     | 11,065                        | 11,449             | 10,575             | 10,135             |
| 1923    | 10,140                    | 11,412                        | 11,517             | 10,788             | 11,015             |
| 1924    | 13,628                    | 12,351                        | 12,787             | 12,499             | 12,816             |
| 1925    | 16,104                    | 13,566                        | 13,740             | 14,759             | 15,386             |
| 1926    | 17,977                    | 15,621                        | 15,166             | 16,627             | 17,918             |
| 1927    | 12,955                    | 13,492                        | 12,692             | 12,678             | 12,842             |
| Average | 12,258                    | 12,460                        | 12,265             | 12,008             | 11,994             |

\* *Yearbook, United States Department of Agriculture, 1928, page 837, Table 250*

† Compiled from office records of the United States Department of Agriculture in Washington.



the years 1915 to 1927 as made by the representatives of that department during four months of the growing season. The table shows also the statistics of final ginnings as officially reported. It will be interesting to observe how close, on an average, the forecasts approximated the actual production.

Further indication of the extent to which the production forecasts approximated the actual production is presented in the following table.

TABLE LXXXV. COMPARISON OF PRODUCTION FORECASTS AND ACTUAL PRODUCTION OF COTTON IN THE UNITED STATES, 1915-1927 <sup>1</sup>

| MONTHS FOR WHICH FORECASTS WERE MADE | PER CENT. THAT AVERAGE OF FORECASTS DIFFERS FROM AVERAGE OF ACTUAL PRODUCTION | COEFFICIENT OF CORRELATION BETWEEN FORECASTS AND ACTUAL PRODUCTION * |
|--------------------------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------|
| July                                 | 1.65                                                                          | .71                                                                  |
| August                               | .06                                                                           | .89                                                                  |
| September                            | -2.04                                                                         | .95                                                                  |
| October                              | -2.14                                                                         | .96                                                                  |

\* Calculated by the sum-product method from data in Table LXXXIV, page 273.

These very close approaches to absolute accuracy in the making of forecasts are striking examples of what may possibly be accomplished by the par method.

### STANDARD ERROR OF ESTIMATE

In any statistical analysis involving the making of estimates or predictions by means of equations formulated from numerical expressions of causal relationship, it is generally desirable to determine in some way the degree of approach to absolute accuracy. Ordinarily, the extent to which estimates and predictions can be made is expected to vary directly in accordance with the magnitude of the coefficient of correlation, though this is not always

<sup>1</sup> Based on data in Table LXXXIV, page 273.



true. If there are abnormally large variables in one series not compensated for in another we may obtain a high expression of relationship which is not a true index of actual cause and effect. This is one of the problems encountered in all methods of correlation, and equations formulated from derived regression coefficients may sometimes be very unsatisfactory for predictive purposes of any kind in spite of high coefficients of correlation. It is only when the coefficient of correlation is an expression of consistent relationship that satisfactory predictive equations can be evolved from the numerical measure of regression.

The error of estimate, as implied, is merely a numerical expression that serves as an indication of the extent to which estimated or predicted values deviate from actual values. It is calculated by extracting the square root of the mean-square-deviation of estimated or predicted values from the actual. Let us turn to Table LXXXII for illustration. As will be seen, the estimated acreages are recorded in column 3, whereas column 2 shows the actual data. The standard error of the estimates in column 3 is calculated by determining the difference between corresponding items in columns 3 and 2, squaring these differences, summing the products, dividing by the number of pairs of items, or the number of differences, and then extracting the square root. The standard error thus calculated may be taken to indicate that if the curve of estimates were normal a distance equal to the standard error measured from either side of the mean of estimates would include approximately 68.26 per cent. of all the estimated values. When the standard error of estimate is larger than the standard deviation of actual values the estimating or predictive equation cannot be relied upon.

When actual changes are correlated with estimated or predicted changes, or when actual values are correlated with estimated or predicted values, it should be possible to obtain a coefficient of correlation which is approximately of the same magnitude as the original coefficient used in formulating the estimating or predictive equation. That is, all or most of the coefficient of correlation should be recovered when actual data are correlated with estimated or predicted values.

### COEFFICIENT OF ALIENATION

Another measure of the degree of approach to accuracy in the making of estimates or predictions is the coefficient of alienation. This measure is based on the mean of squared deviations of actual items from their arithmetic mean and on the mean of squared deviations of estimated or predicted values from the actual values, and it involves the calculation of both the standard error of estimate and the standard deviation. For facilitating its calculation the following formula may be found convenient:

$$ca = \frac{SEy^2}{SDy^2},$$

in which

$ca$  = the coefficient of alienation

$SEy^2$  = the square of the standard error of estimate

$SDy^2$  = the square of the standard deviation of actual values.

Since the standard deviation is merely the root-mean-square of deviations from the arithmetic mean of the actual data, and since the standard error of estimate is simply the root-mean-square of the differences between actual and estimated or predicted values, it necessarily follows that when the square of the standard error is

divided by the square of the standard deviation a quotient is obtained that represents the root-mean-square of the differences between actual and estimated or predicted values in terms of the root-mean-square deviations of actual values from their arithmetic mean. Merely dividing the standard error by the standard deviation would give a quotient of little or no significance. This is because both the standard error and the standard deviation are derived from the means of squared deviations, and not simply from the deviations as such. It is no more than a truism to state that the squares of numbers do not increase in the same proportion as the magnitudes of the numbers themselves. The square of 12, for example, is 144, whereas the square of 17 is 289, and 289 is more than twice the size of 144, although 17 is less than one and a half times the size of 12.

## CHAPTER XV

### INDEX NUMBERS

The construction of index numbers affords one of the best means for expressing numerically the relative magnitudes of changing or changed conditions from time to time. If it is desired, for example, to show the trend of the motor vehicle industry in the United States for a period, say, of ten years, it may be of greater significance to express the number of vehicles produced in any one year as a percentage of the number produced in a fixed base year than to present only the absolute statistics of actual production. This is particularly true if interest lies only in the relation of quantity production in any one year to quantity production in a certain other designated year. The principal reason an index number serves so admirably as a means of measuring or expressing relative magnitudes is that, being expressed as a percentage, its significance is more readily grasped than that of large absolute numbers.

As the name implies, an index number is simply a number that represents relative changes. It shows the magnitude of one number or group of numbers in relation to some other similar number or group of numbers. It is in this sense, therefore, that an index number may be considered as representing comparative magnitudes. A brief inquiry will be made into the two common types of index numbers.

## SIMPLE INDEX NUMBERS

A simple index number is constructed to show the changes in a single variable. If we have the annual statistics of bank deposits in the United States for a period of years and wish to show relative changes it is only necessary to select a base year, divide the deposits for that year into the deposits for each respective year, and multiply by 100. Any value thus obtained is the deposits of any specified year expressed as a percentage of the base year deposits, and is, therefore, an index. The following table will simplify its calculation.

TABLE LXXXVI. BANK DEPOSITS IN THE UNITED STATES, 1921-1927 <sup>1</sup>  
(Simple Index Number)

| YEAR | TOTAL DEPOSITS<br><i>1,000,000 Dollars</i> | INDEX NUMBER OF DEPOSITS<br>(1921 as Base) |
|------|--------------------------------------------|--------------------------------------------|
| 1921 | 34,845                                     | 100                                        |
| 1922 | 37,194                                     | 107                                        |
| 1923 | 40,034                                     | 115                                        |
| 1924 | 42,954                                     | 123                                        |
| 1925 | 46,766                                     | 134                                        |
| 1926 | 48,882                                     | 140                                        |
| 1927 | 51,133                                     | 147                                        |

The index numbers in the table represent the deposits of specified years as percentages of deposits for 1921, the base year. Any base year other than 1921 could just as well be used, and even an average for two or more years could be employed as the basic value.

There is quite often no distinction made between relative numbers and index numbers, but some consider it a better practice to employ the term "relative numbers" when referring to the expression of the items in a single

<sup>1</sup> *Statistical Abstract of the United States, 1928*, page 258, Table 262.

series in terms of a fixed base. By such distinction, the term "index numbers" is then reserved for use when two or more series are combined and expressed in terms of a fixed base. The validity of this practice is doubtful.

### LINK INDEX NUMBERS

If it is desired to compare values for each year, such as prices or production, with those of the immediately preceding year a link relative index number may be found very satisfactory. It is constructed by letting each year be a base year and then expressing each succeeding year in terms of the respective base. Accordingly, in Table LXXXVI the value given for 1921 would be divided into the 1922 value and the quotient multiplied by 100 to obtain the link relative for 1922. The 1923 link relative index number would be obtained by dividing the 1923 value by the 1922 value and then multiplying by 100.

### COMPOSITE INDEX NUMBERS

Whenever it is desired to express the changes in two or more conditions considered collectively we construct a composite index number, so-called because it is composed of two or more elements or parts. The method of constructing such an index will depend largely upon the character of the data and the purpose for which the number is to be used.

A composite index number may be either unweighted or weighted. In the unweighted index the values for the base year are summated, the total divided into itself, and the quotient multiplied by 100 to obtain the unweighted composite index number for that year. The total for the base year is then divided into the total for each of the other years and the quotient multiplied



by 100 to obtain the corresponding index numbers. This method of calculation is illustrated in Table LXXXVII.<sup>1</sup>

TABLE LXXXVII. YIELD OF FLAXSEED PER ACRE IN MINNESOTA AND NORTH DAKOTA, 1921-1927 <sup>2</sup>

(Unweighted Composite Index Number)

| YEAR | YIELD PER ACRE |                | UNWEIGHTED COMPOSITE<br>INDEX NUMBER OF YIELD<br>PER ACRE<br>(1921 as Base) |
|------|----------------|----------------|-----------------------------------------------------------------------------|
|      | Minnesota      | North Dakota   |                                                                             |
|      | <i>Bushels</i> | <i>Bushels</i> |                                                                             |
| 1921 | 9.5            | 6.5            | 100                                                                         |
| 1922 | 10.0           | 9.3            | 121                                                                         |
| 1923 | 10.0           | 7.7            | 111                                                                         |
| 1924 | 11.4           | 8.5            | 124                                                                         |
| 1925 | 10.0           | 6.5            | 103                                                                         |
| 1926 | 9.4            | 5.5            | 93                                                                          |
| 1927 | 9.7            | 8.2            | 112                                                                         |

By this method the unweighted index number of yield is calculated by summing the yields for the two states, dividing the total for each year by the total for the base year, and then multiplying by 100. Obviously, this gives an unweighted index of yield because no adjustment is made for differences in acreage and production in the two states.

Using the data on flaxseed, let us now proceed to the calculation of the weighted index of yield per acre. This will be seen to be very simple, since the weighted average should already be well understood. The following table illustrates the principles involved in calculation.

<sup>1</sup> A composite index number may be constructed also by averaging the so-called "relatives." By this method the items in each series are expressed as percentages of a fixed base, the percentages summated, and then divided by the number to obtain the desired index.

<sup>2</sup> Yearbooks, *United States Department of Agriculture*, 1928, page 739, Table 94, and 1926, page 871, Table 89.

TABLE LXXXVIII. ACREAGE AND PRODUCTION OF FLAXSEED IN MINNESOTA AND NORTH DAKOTA, 1921-1927 <sup>1</sup>

(Weighted Composite Index Number of Yield per Acre)

| YEAR | ACREAGE                |                        | PRODUCTION               |                          | WEIGHTED<br>COMPOSITE<br>INDEX NUMBER<br>OF YIELD<br>PER ACRE<br>(1921 as Base) |
|------|------------------------|------------------------|--------------------------|--------------------------|---------------------------------------------------------------------------------|
|      | Minnesota              | North Dakota           | Minnesota                | North Dakota             |                                                                                 |
|      | <i>1,000<br/>Acres</i> | <i>1,000<br/>Acres</i> | <i>1,000<br/>Bushels</i> | <i>1,000<br/>Bushels</i> |                                                                                 |
| 1921 | 314                    | 430                    | 2,983                    | 2,795                    | 100                                                                             |
| 1922 | 310                    | 521                    | 3,100                    | 4,845                    | 123                                                                             |
| 1923 | 527                    | 1,050                  | 5,270                    | 8,050                    | 108                                                                             |
| 1924 | 712                    | 1,873                  | 8,117                    | 15,920                   | 119                                                                             |
| 1925 | 740                    | 1,461                  | 7,400                    | 9,496                    | 99                                                                              |
| 1926 | 814                    | 1,380                  | 7,652                    | 7,590                    | 88                                                                              |
| 1927 | 757                    | 1,242                  | 7,343                    | 10,184                   | 113                                                                             |

In calculating the weighted indices the weighted average yield for the base year, 1921, is divided into the weighted average yield for each of the years and the quotient multiplied by 100. By this method of calculation the acreage and production in each of the two states are given weights according to their relative magnitudes, so that a relatively high or low yield in one state is not stressed beyond its own importance.

The weighting of index numbers has many applications. One of the most common of these is in the study of changes in cost of living. When an attempt is being made to show the relation between costs of living at different times, the analyst is interested not only in the items of unit cost, but also in the relation between total costs of each of the items entering into the cost of living. He wishes to know, for example, what part the cost of meat constitutes of the total of all the individual costs. To facilitate the calcula-

<sup>1</sup> Yearbooks, United States Department of Agriculture, 1928, page 739, Table 93; 1926, page 871, Table 88; and 1923, page 708, Table 150.

tion of a weighted cost of living index number, each commodity entering into this cost is given weight according to its own importance in relation to the total. Each of these weights is then multiplied by the unit cost of the particular commodity, the products summated, divided by the total of corresponding costs for the base year, and then multiplied by 100 to obtain the weighted index for the specified year. It is obvious that by giving weight in this manner to each individual commodity the relative importance of high or low unit costs is not over-emphasized, nor is it minimized.

#### ACCURACY OF INDEX NUMBERS

Much has been written as to the factors that contribute to the accuracy and dependability of index numbers. Perhaps the most outstanding analysis of this phase of the subject has been made by the renowned economist, Professor Irving Fisher. In his famous work, *The Making of Index Numbers*, he points out two tests which may be applied in measuring accuracy and dependability. The first of these is the *time-reversal* test. When a change in the base does not affect the relative sizes of the index numbers they are said to meet the requirements of the *time-reversal* test. That is to say, if in constructing index numbers of wholesale prices of certain farm products with 1912 as the base it is found that the index for 1912 is one-half as great as for 1919, then changing the base to 1919 should not alter the relative magnitude of these indices. In other words, the 1919 index should be twice the size of the 1912 index regardless of which of the two years is selected as a base. Obviously, this is a logical test to make, but in actual practice it is doubtful if any composite index number could be expected to meet it. Among many in-

stances that may be cited, this is particularly true of a cost of living index number in the construction of which the weights remain constant, or nearly so, whereas relative price changes may vary considerably from year to year.

The second of the two tests referred to is the *factor-reversal* test. Its significance will be immediately appreciated by referring to quantity of cotton marketed and price received for it. Let us say, with the year 1927 as the base, that the index of quantity of cotton sold on a certain day in 1924 at the New Orleans market is 105, and that the index of price is 115. If the index of price meets the *factor-reversal* test the product of the quantity and price indices, 105 and 115, should be equal to the index of value. ( $1.05 \times 1.15 \times 100$ ).

The making of index numbers involves many ramifications which are beyond the scope of this text. Only a discussion of the more elementary principles can be attempted here, with the aim to give the student a definite idea as to the nature of indices and the principal factors to be considered in their construction. In more advanced studies it will be found profitable to take recourse to some of the more detailed and comprehensive works devoted especially to index numbers.<sup>1</sup>

<sup>1</sup> See Fisher, *The Making of Index Numbers*, Mitchell, *Index Numbers of Wholesale Prices in the United States*; Jerome, *Statistical Method*, edition of 1924, Chapters XI and XII; Day, *Statistical Analysis*, edition of 1927, Chapters XXI, XXII, and XXIII; Bureau of Labor Statistics, *Index Numbers of Wholesale Prices*; and National Fertilizer Association, *The Weekly Wholesale Price Index*.

## APPENDICES





## APPENDIX A

### ORDINATES OF THE PROBABILITY CURVE

In Table LXXXIX are recorded the ordinates of the probability curve in terms of percentages of the mean ordinate that are associated with varying distances on the x axis as measured from the mean in number of standard deviations. For example, when a distance of 1 standard deviation is measured from either side of the mean in the normal probability curve the corresponding ordinate on the y axis is .60653 of the ordinate of the mean. A distance of 2 standard deviations measured from the mean is associated with a y ordinate that is .13534 as great as the ordinate of the mean, whereas a distance from the mean of 1.55 standard deviations is associated with a y ordinate that is .30082 as great in magnitude as the ordinate with which the mean coincides.

It is interesting to note the way by which a curve of probability may be constructed from the values in the table. Let us assume, for instance, that we have a mean of 75, a standard deviation of 12, and an x value of 87. To determine the y ordinate for the x value we merely subtract 75, the mean, from 87 and then divide by 12, the standard deviation. As will be seen, 87 minus 75 gives 12, which divided by the standard deviation gives a value of 1. This means that the ordinate of the probability curve is to be plotted on the y axis at a point coinciding with a linear distance of 1 standard deviation measured from the average on the x axis. By referring to Table LXXXIX it has already been seen that an x value of 1 standard deviation measured from the mean of a probability curve is associated with a y value that is .60653 as great in size as the y ordinate of the mean. It is obvious, therefore, that the x value of 87 will coincide with a y ordinate that is .60653 the size of the ordinate of the mean. Other y ordinates of the probability curve would be determined in like manner.

TABLE LXXXIX. ORDINATES OF THE PROBABILITY CURVE<sup>1</sup>

| Sigma<br>(Standard Deviation) | .00     | .01    | .02    | .03    | .04    | .05    | .06    | .07    | .08    | .09    |
|-------------------------------|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0                           | 1.00000 | .99995 | .99980 | .99955 | .99920 | .99875 | .99820 | .99755 | .99680 | .99596 |
| 0.1                           | .99501  | .99397 | .99283 | .99159 | .99025 | .98881 | .98728 | .98565 | .98393 | .98211 |
| 0.2                           | .98020  | .97819 | .97609 | .97390 | .97161 | .96923 | .96676 | .96421 | .96156 | .95882 |
| 0.3                           | .95600  | .95309 | .95009 | .94701 | .94384 | .94059 | .93725 | .93384 | .93034 | .92677 |
| 0.4                           | .92312  | .91939 | .91558 | .91169 | .90774 | .90371 | .89960 | .89543 | .89119 | .88688 |
| 0.5                           | .88250  | .87805 | .87354 | .86897 | .86433 | .85963 | .85487 | .85006 | .84518 | .84025 |
| 0.6                           | .83527  | .83023 | .82514 | .82000 | .81481 | .80957 | .80429 | .79896 | .79358 | .78816 |
| 0.7                           | .78270  | .77721 | .77167 | .76609 | .76048 | .75484 | .74916 | .74345 | .73771 | .73194 |
| 0.8                           | .72615  | .72033 | .71448 | .70861 | .70272 | .69680 | .69087 | .68492 | .67896 | .67297 |
| 0.9                           | .66698  | .66097 | .65495 | .64892 | .64288 | .63683 | .63078 | .62472 | .61866 | .61259 |
| 1.0                           | .60653  | .60047 | .59440 | .58834 | .58228 | .57623 | .57018 | .56414 | .55811 | .55209 |
| 1.1                           | .54607  | .54007 | .53409 | .52811 | .52215 | .51620 | .51028 | .50437 | .49848 | .49260 |
| 1.2                           | .48675  | .48092 | .47511 | .46933 | .46357 | .45783 | .45212 | .44644 | .44078 | .43516 |
| 1.3                           | .42956  | .42399 | .41845 | .41294 | .40747 | .40202 | .39661 | .39123 | .38589 | .38058 |
| 1.4                           | .37531  | .37007 | .36488 | .35971 | .35459 | .34950 | .34445 | .33944 | .33447 | .32954 |
| 1.5                           | .32465  | .31980 | .31500 | .31023 | .30550 | .30082 | .29618 | .29158 | .28702 | .28251 |
| 1.6                           | .27804  | .27361 | .26923 | .26489 | .26059 | .25634 | .25213 | .24797 | .24385 | .23978 |
| 1.7                           | .23575  | .23176 | .22782 | .22392 | .22007 | .21627 | .21250 | .20879 | .20511 | .20148 |
| 1.8                           | .19790  | .19436 | .19086 | .18741 | .18400 | .18064 | .17732 | .17404 | .17081 | .16762 |
| 1.9                           | .16447  | .16137 | .15831 | .15529 | .15232 | .14938 | .14649 | .14364 | .14083 | .13806 |

<sup>1</sup> Tables for Statisticians and Biometricians, Cambridge University, by W. F. Sheppard; edited by Karl Pearson.



## AREAS OF THE PROBABILITY SURFACE

In the study of probabilities, or of normal frequency distributions, it is of interest to know what portion of the frequency surface lies within specified distances measured from the arithmetic mean. The study of probability has already shown the mean to be the item coinciding with the midpoint of the distribution, with an equal number of items both below and above it. It will be recalled also that in a normal frequency distribution a distance equal to the standard deviation when measured from either side of the mean will include an area embracing approximately 68.26 per cent. of the total number of items. This can be read directly from Table XC, where, it is seen, to be exact, that a distance of 1 standard deviation measured off either to the right or left of the mean includes .34134 of the total surface. A distance of .06 standard deviation measured from one side of the mean includes .02392 of all the items, and a distance of 1.45 standard deviations thus measured embraces an area including .42647 of all the items. It is obvious, of course, that when these distances are measured on the  $x$  axis from both sides of the mean the area embraced will include twice the portion of the surface included when measurement is made from one side only.

For illustration let us assume a mean of 75, a single value of 96, and a standard deviation of 12. To determine the portion of the entire probability surface lying between the mean and the value of 96 it is but necessary to subtract 75 from 96, divide by the standard deviation, and then read from the table the value corresponding to the quotient. The difference between 96 and 75 is 21. Dividing this difference by 12, the standard deviation, gives 1.75. A distance of 1.75 standard deviations measured from the mean will be seen to include .45994 of the total surface, or .91988 of the total surface when measured from both sides of the mean. Thus it is that the area lying between the mean and any other value can be determined.

TABLE XC. AREAS OF THE PROBABILITY SURFACE <sup>1</sup>

| Sigma<br>(Standard Deviation) | .00    | .01    | .02    | .03    | .04    | .05    | .06    | .07    | .08    | .09    |
|-------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0                           | .00000 | .00399 | .00798 | .01197 | .01595 | .01994 | .02392 | .02790 | .03188 | .03586 |
| 0.1                           | .03983 | .04380 | .04776 | .05172 | .05567 | .05962 | .06356 | .06749 | .07142 | .07535 |
| 0.2                           | .07926 | .08317 | .08706 | .09095 | .09483 | .09871 | .10257 | .10642 | .11026 | .11409 |
| 0.3                           | .11791 | .12172 | .12552 | .12930 | .13307 | .13683 | .14058 | .14431 | .14803 | .15173 |
| 0.4                           | .15542 | .15910 | .16276 | .16640 | .17003 | .17364 | .17724 | .18082 | .18439 | .18793 |
| 0.5                           | .19146 | .19497 | .19847 | .20194 | .20540 | .20884 | .21226 | .21566 | .21904 | .22240 |
| 0.6                           | .22575 | .22907 | .23237 | .23565 | .23891 | .24215 | .24537 | .24857 | .25175 | .25490 |
| 0.7                           | .25804 | .26115 | .26424 | .26730 | .27035 | .27337 | .27637 | .27935 | .28230 | .28524 |
| 0.8                           | .28814 | .29103 | .29389 | .29673 | .29955 | .30234 | .30511 | .30785 | .31057 | .31327 |
| 0.9                           | .31594 | .31859 | .32121 | .32381 | .32639 | .32894 | .33147 | .33398 | .33646 | .33891 |
| 1.0                           | .34134 | .34375 | .34614 | .34850 | .35083 | .35314 | .35543 | .35769 | .35993 | .36214 |
| 1.1                           | .36433 | .36650 | .36864 | .37076 | .37286 | .37493 | .37689 | .37900 | .38100 | .38298 |
| 1.2                           | .38493 | .38686 | .38877 | .39065 | .39251 | .39435 | .39617 | .39796 | .39973 | .40147 |
| 1.3                           | .40320 | .40490 | .40658 | .40824 | .40988 | .41149 | .41309 | .41466 | .41621 | .41774 |
| 1.4                           | .41924 | .42073 | .42220 | .42364 | .42507 | .42647 | .42786 | .42922 | .43056 | .43198 |
| 1.5                           | .43319 | .43448 | .43574 | .43699 | .43822 | .43943 | .44062 | .44179 | .44295 | .44408 |
| 1.6                           | .44520 | .44630 | .44738 | .44845 | .44950 | .45053 | .45154 | .45254 | .45352 | .45449 |
| 1.7                           | .45543 | .45637 | .45728 | .45818 | .45907 | .45994 | .46080 | .46164 | .46246 | .46327 |
| 1.8                           | .46047 | .46485 | .46562 | .46638 | .46712 | .46784 | .46856 | .46926 | .46995 | .47062 |
| 1.9                           | .47128 | .47193 | .47257 | .47320 | .47381 | .47441 | .47500 | .47558 | .47615 | .47670 |

<sup>1</sup> Tables for Statisticians and Biometricians, Cambridge University, by W. F. Sheppard; edited by Karl Pearson.





## APPENDIX B

### LOGARITHMS <sup>1</sup>

The logarithm of a number is the exponent of the power to which a fixed number, known as the base, must be raised in order to equal or produce a certain given number. There are two systems of logarithms in general use, the Naperian system, with 2.71827 as the base, which is used extensively in analytical work, and the common, or Briggsian, system, with 10 as the base, which is used in practical computations.

The principal value of logarithms is that they may be used to reduce the time and work in multiplication and division, and in extracting the roots of numbers. They are also useful in computing the powers of numbers.

Logarithms are made up of two parts, the characteristic and the mantissa. The characteristic is the integral part of a logarithm, or the functional part that remains constant. The mantissa is the part or portion of a logarithm that is expressed as a decimal. For example, the logarithm of 400 is 2.6021, in which the characteristic is 2 and the mantissa .6021. An important feature of logarithms is that the characteristic is determined by that part of the number lying to the left of the decimal, its size being one less than the number of digits preceding the decimal point. Thus, the characteristic of the logarithms of numbers from 1 to 9, inclusive, is 0. In like manner, the characteristic of logarithms of numbers from 10 to 99 is 1, whereas the characteristic of logarithms of numbers from 100 to 999 is 2. The characteristic of logarithms of numbers from 1,000 to 9,999 is 3, and so on for all other numbers. Another important feature of logarithms is that the mantissa, the fractional portion, does not depend upon the position of the decimal point in the number. It will be seen, therefore, that while the characteristic of the logarithms of 5, 50, and 500 are 0, 1, and 2, respectively, the mantissa is .6990 for

<sup>1</sup> Adapted from *Accountants' Handbook*, The Ronald Press Company, 1923, and other sources.

all of them. The mantissa may be expressed with a large number of digits, but for practical purposes a four-place fraction will suffice, the last digit being rounded off in the same way that 96.578 per cent. is expressed as 96.58 or as 96.6 per cent.

Logarithms of numbers are found by first determining the characteristic and then looking up the mantissa. Suppose we wish to know the logarithm of 54.5. By the general rule that has already been outlined, the characteristic is seen to be 1. The mantissa is found by turning to the table of logarithms and locating the fraction corresponding to 54.5. This fraction is seen to be .7364. The logarithm of 54.5, therefore, is 1.7364. If the number were 54.55 instead of 54.5 it would be necessary to interpolate for the difference, unless the table shows logarithms for such numbers. This is done by determining the difference between the mantissa of the logarithm of 54.5 and the logarithm of 54.6, multiplying this difference by .5, and adding the product to the mantissa of the logarithm of 54.5. Accordingly, as we have seen, the mantissa of the logarithm of 54.5 is .7364, while the mantissa of the logarithm of 54.6 is .7372. The difference between them is .0008. Multiplying .0008 by .5 gives .00040, which when added to .7364, the mantissa of the logarithm of 54.5, gives .7368, the mantissa of the logarithm of 54.55. If the number were 54.53 the difference of .0008 would be multiplied by .3 to obtain the product to be added to .7364 to give the correct mantissa, and so on for any other number.

To find the number corresponding to a given logarithm, it is but necessary to locate the mantissa of the logarithm in the table and then read off the number corresponding thereto. Suppose the logarithm is 1.9053. Reading down the table of logarithms we find the fraction .9053 is located opposite 8.0 in the column headed 4. The characteristic, 1, tells us that there must be two digits in that part of the number preceding the decimal. Accordingly, we record 80 as the whole number and .4 as the fraction. The number corresponding to the logarithm, therefore, is 80.4. If the logarithm were 1.9056 the number corresponding thereto could be determined by interpolating for the difference.

To multiply, or to find the product of two numbers, we merely look up their logarithms, summate these logarithms, look up the number corresponding to the mantissa, or fractional part

of the sum of the two logarithms, and then move the decimal point to the right as many places as it takes to give the product one more digit than the size of the characteristic of the sum of the two logarithms. For example, let us illustrate how the product of 22 and 26 is obtained by the use of logarithms. The logarithm of 22 is 1.3424, and of 26 it is 1.4150. The summation of these two logarithms is 2.7574. The number corresponding to .7574, the fractional part of the summation, is 572. This number, then, 572, is the product of 22 and 26.

The root of a number may be extracted by dividing the logarithm of the number by the number indicating the root to be found, and then finding in the table the antilogarithm of the quotient. For example, in extracting the square root of 144, we first look up the logarithm of 144, which is 2.1584. Dividing 2.1584 by 2, the number indicating the root to be found, gives a quotient of 1.0792. By referring to the table of logarithms, the number corresponding to .0792 is found to be 12, which is the square root of 144. The cube root of 144 is extracted by dividing its logarithm by 3 and then referring to the table of logarithms for the number corresponding to the quotient. We have, therefore, the logarithm of 144, which is 2.1584, divided by 3, giving a quotient of approximately .7195. The number corresponding to .7195, therefore, is the cube root of 144. Any other root is extracted by the same method.

In division, the difference between the logarithms of the dividend and divisor are determined by subtraction and then the quotient obtained by looking up the number corresponding to this difference between the two logarithms. For example, we wish to divide 144 by 12. Turning to the table we find the logarithms of 144 and 12 to be 2.1584 and 1.0792, respectively. Subtracting 1.0792, the logarithm of the divisor, from 2.1584, the logarithm of the dividend, leaves a difference of 1.0792. Looking up the number corresponding to .0792 in the table, we find it to be 12, the quotient of 144 divided by 12.

If the dividend is smaller than the divisor, so that the quotient is less than unity, another step must be included in the process of division by the use of logarithms. An illustration will make this easily understood. Suppose we wish to divide 6 by 7. Turning to the table we find the logarithm of 6 to be .7782, whereas

the logarithm of 7 is .8451. In subtracting .8451 from .7782 we first prefix 10, as follows, and then subtract:

$$\begin{array}{r} 10 \quad .7782 \\ \quad .8451 \\ \hline 9 \quad .9331 \end{array}$$

Turning again to the table we find the number .857 as the value of the fraction .9331. The quotient of 6 divided by 7, therefore, is .857.

TABLE XCI. LOGARITHMS <sup>1</sup>

| No. | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1.0 | .0000 | .0043 | .0086 | .0128 | .0170 | .0212 | .0253 | .0294 | .0334 | .0374 |
| .1  | .0414 | .0453 | .0492 | .0531 | .0569 | .0607 | .0645 | .0682 | .0719 | .0755 |
| .2  | .0792 | .0828 | .0864 | .0899 | .0934 | .0969 | .1004 | .1038 | .1072 | .1106 |
| .3  | .1139 | .1173 | .1206 | .1239 | .1271 | .1303 | .1335 | .1367 | .1399 | .1430 |
| .4  | .1461 | .1492 | .1523 | .1553 | .1584 | .1614 | .1644 | .1673 | .1703 | .1732 |
| .5  | .1761 | .1790 | .1818 | .1847 | .1875 | .1903 | .1931 | .1959 | .1987 | .2014 |
| .6  | .2041 | .2068 | .2095 | .2122 | .2148 | .2175 | .2201 | .2227 | .2253 | .2279 |
| .7  | .2304 | .2330 | .2355 | .2380 | .2405 | .2430 | .2455 | .2480 | .2504 | .2529 |
| .8  | .2553 | .2577 | .2601 | .2625 | .2648 | .2672 | .2695 | .2718 | .2742 | .2765 |
| .9  | .2788 | .2810 | .2833 | .2856 | .2878 | .2900 | .2923 | .2945 | .2967 | .2989 |
| 2.0 | .3010 | .3032 | .3054 | .3075 | .3096 | .3118 | .3139 | .3160 | .3181 | .3201 |
| .1  | .3222 | .3243 | .3263 | .3284 | .3304 | .3324 | .3345 | .3365 | .3385 | .3404 |
| .2  | .3424 | .3444 | .3464 | .3483 | .3502 | .3522 | .3541 | .3560 | .3579 | .3598 |
| .3  | .3617 | .3636 | .3655 | .3674 | .3692 | .3711 | .3729 | .3747 | .3766 | .3784 |
| .4  | .3802 | .3820 | .3838 | .3856 | .3874 | .3892 | .3909 | .3927 | .3945 | .3962 |
| .5  | .3979 | .3997 | .4014 | .4031 | .4048 | .4065 | .4082 | .4099 | .4116 | .4133 |
| .6  | .4150 | .4166 | .4183 | .4200 | .4216 | .4232 | .4249 | .4265 | .4281 | .4298 |
| .7  | .4314 | .4330 | .4346 | .4362 | .4378 | .4393 | .4409 | .4425 | .4440 | .4450 |
| .8  | .4472 | .4487 | .4502 | .4518 | .4533 | .4548 | .4564 | .4579 | .4594 | .4609 |
| .9  | .4624 | .4639 | .4654 | .4669 | .4683 | .4698 | .4713 | .4728 | .4742 | .4757 |
| 3.0 | .4771 | .4786 | .4800 | .4814 | .4829 | .4843 | .4857 | .4871 | .4886 | .4900 |
| .1  | .4914 | .4928 | .4942 | .4955 | .4969 | .4983 | .4997 | .5011 | .5024 | .5038 |
| .2  | .5051 | .5065 | .5079 | .5092 | .5105 | .5119 | .5132 | .5145 | .5159 | .5172 |
| .3  | .5185 | .5198 | .5211 | .5224 | .5237 | .5250 | .5263 | .5276 | .5289 | .5302 |
| .4  | .5315 | .5328 | .5340 | .5353 | .5366 | .5378 | .5391 | .5403 | .5416 | .5428 |
| .5  | .5441 | .5453 | .5465 | .5478 | .5490 | .5502 | .5514 | .5527 | .5539 | .5551 |
| .6  | .5563 | .5575 | .5587 | .5599 | .5611 | .5623 | .5635 | .5647 | .5658 | .5670 |
| .7  | .5682 | .5694 | .5705 | .5717 | .5729 | .5740 | .5752 | .5763 | .5775 | .5786 |
| .8  | .5798 | .5809 | .5821 | .5832 | .5843 | .5855 | .5866 | .5877 | .5888 | .5899 |
| .9  | .5911 | .5922 | .5933 | .5944 | .5955 | .5966 | .5977 | .5988 | .5999 | .6010 |
| 4.0 | .6021 | .6031 | .6042 | .6053 | .6064 | .6075 | .6085 | .6096 | .6107 | .6117 |
| .1  | .6128 | .6138 | .6149 | .6160 | .6170 | .6180 | .6194 | .6201 | .6212 | .6222 |
| .2  | .6232 | .6243 | .6253 | .6263 | .6274 | .6284 | .6294 | .6304 | .6314 | .6325 |
| .3  | .6335 | .6345 | .6355 | .6365 | .6375 | .6385 | .6395 | .6405 | .6415 | .6425 |
| .4  | .6435 | .6444 | .6454 | .6464 | .6474 | .6484 | .6493 | .6503 | .6513 | .6522 |
| .5  | .6532 | .6542 | .6551 | .6561 | .6571 | .6580 | .6590 | .6599 | .6609 | .6618 |
| .6  | .6628 | .6637 | .6646 | .6656 | .6665 | .6675 | .6684 | .6693 | .6702 | .6712 |
| .7  | .6721 | .6730 | .6739 | .6749 | .6758 | .6767 | .6776 | .6785 | .6794 | .6803 |
| .8  | .6812 | .6821 | .6830 | .6839 | .6848 | .6857 | .6866 | .6875 | .6884 | .6893 |
| .9  | .6902 | .6911 | .6920 | .6928 | .6937 | .6946 | .6955 | .6964 | .6972 | .6981 |
| 5.0 | .6990 | .6998 | .7007 | .7016 | .7024 | .7033 | .7042 | .7050 | .7059 | .7067 |
| .1  | .7076 | .7084 | .7093 | .7101 | .7110 | .7118 | .7126 | .7135 | .7143 | .7152 |
| .2  | .7160 | .7168 | .7177 | .7185 | .7193 | .7202 | .7210 | .7218 | .7226 | .7235 |
| .3  | .7243 | .7251 | .7259 | .7267 | .7275 | .7284 | .7292 | .7300 | .7308 | .7316 |
| .4  | .7324 | .7332 | .7340 | .7348 | .7356 | .7364 | .7372 | .7380 | .7388 | .7396 |

<sup>1</sup> There are many books containing logarithms that have been published, and which may be found very convenient. Special attention is called to the *Accountants' Handbook*, 1923, edited by Earl A. Saliers, containing five-place



TABLE XCI. LOGARITHMS—*Continued*

| No. | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| .5  | .7404 | .7412 | .7419 | .7427 | .7435 | .7443 | .7451 | .7459 | .7466 | .7474 |
| .6  | .7482 | .7490 | .7497 | .7505 | .7513 | .7520 | .7528 | .7536 | .7543 | .7551 |
| .7  | .7559 | .7566 | .7574 | .7582 | .7589 | .7597 | .7604 | .7612 | .7619 | .7627 |
| .8  | .7634 | .7642 | .7649 | .7657 | .7664 | .7672 | .7679 | .7686 | .7694 | .7701 |
| .9  | .7709 | .7716 | .7723 | .7731 | .7738 | .7745 | .7752 | .7760 | .7767 | .7774 |
| 6.0 | .7782 | .7789 | .7796 | .7803 | .7810 | .7818 | .7825 | .7832 | .7839 | .7846 |
| .1  | .7853 | .7860 | .7868 | .7875 | .7882 | .7889 | .7896 | .7903 | .7910 | .7917 |
| .2  | .7924 | .7931 | .7938 | .7945 | .7952 | .7959 | .7966 | .7973 | .7980 | .7987 |
| .3  | .7993 | .8000 | .8007 | .8014 | .8021 | .8028 | .8035 | .8041 | .8048 | .8055 |
| .4  | .8062 | .8069 | .8075 | .8082 | .8089 | .8096 | .8102 | .8109 | .8116 | .8122 |
| .5  | .8129 | .8136 | .8142 | .8149 | .8156 | .8162 | .8169 | .8176 | .8182 | .8189 |
| .6  | .8195 | .8202 | .8209 | .8215 | .8222 | .8228 | .8235 | .8241 | .8248 | .8254 |
| .7  | .8261 | .8267 | .8274 | .8280 | .8287 | .8293 | .8299 | .8306 | .8312 | .8319 |
| .8  | .8325 | .8331 | .8338 | .8344 | .8351 | .8357 | .8363 | .8370 | .8376 | .8382 |
| .9  | .8388 | .8395 | .8401 | .8407 | .8414 | .8420 | .8426 | .8432 | .8439 | .8445 |
| 7.0 | .8451 | .8457 | .8463 | .8470 | .8476 | .8482 | .8488 | .8494 | .8500 | .8506 |
| .1  | .8513 | .8519 | .8525 | .8531 | .8537 | .8543 | .8549 | .8555 | .8561 | .8567 |
| .2  | .8573 | .8579 | .8585 | .8591 | .8597 | .8603 | .8609 | .8615 | .8621 | .8627 |
| .3  | .8633 | .8639 | .8645 | .8651 | .8657 | .8663 | .8669 | .8675 | .8681 | .8686 |
| .4  | .8692 | .8698 | .8704 | .8710 | .8716 | .8722 | .8727 | .8733 | .8739 | .8745 |
| .5  | .8751 | .8756 | .8762 | .8768 | .8774 | .8779 | .8785 | .8791 | .8797 | .8802 |
| .6  | .8808 | .8814 | .8820 | .8825 | .8831 | .8837 | .8842 | .8848 | .8854 | .8859 |
| .7  | .8865 | .8871 | .8876 | .8882 | .8887 | .8893 | .8899 | .8904 | .8910 | .8915 |
| .8  | .8921 | .8927 | .8932 | .8938 | .8943 | .8949 | .8954 | .8960 | .8965 | .8971 |
| .9  | .8976 | .8982 | .8987 | .8993 | .8998 | .9004 | .9009 | .9015 | .9020 | .9025 |
| 8.0 | .9031 | .9036 | .9042 | .9047 | .9053 | .9058 | .9063 | .9069 | .9074 | .9079 |
| .1  | .9085 | .9090 | .9096 | .9101 | .9106 | .9112 | .9117 | .9122 | .9128 | .9133 |
| .2  | .9138 | .9143 | .9149 | .9154 | .9159 | .9165 | .9170 | .9175 | .9180 | .9186 |
| .3  | .9191 | .9196 | .9201 | .9206 | .9212 | .9217 | .9222 | .9227 | .9232 | .9238 |
| .4  | .9243 | .9248 | .9253 | .9258 | .9263 | .9269 | .9274 | .9279 | .9284 | .9289 |
| .5  | .9294 | .9299 | .9304 | .9309 | .9315 | .9320 | .9325 | .9330 | .9335 | .9340 |
| .6  | .9345 | .9350 | .9355 | .9360 | .9365 | .9370 | .9375 | .9380 | .9385 | .9390 |
| .7  | .9395 | .9400 | .9405 | .9410 | .9415 | .9420 | .9425 | .9430 | .9435 | .9440 |
| .8  | .9445 | .9450 | .9455 | .9460 | .9465 | .9469 | .9474 | .9479 | .9484 | .9489 |
| .9  | .9494 | .9499 | .9504 | .9509 | .9513 | .9518 | .9523 | .9528 | .9533 | .9538 |
| 9.0 | .9542 | .9547 | .9552 | .9557 | .9562 | .9566 | .9571 | .9576 | .9581 | .9586 |
| .1  | .9590 | .9595 | .9600 | .9605 | .9609 | .9614 | .9619 | .9624 | .9628 | .9633 |
| .2  | .9638 | .9643 | .9647 | .9652 | .9657 | .9661 | .9666 | .9671 | .9675 | .9680 |
| .3  | .9685 | .9689 | .9694 | .9699 | .9703 | .9708 | .9713 | .9717 | .9722 | .9727 |
| .4  | .9731 | .9736 | .9741 | .9745 | .9750 | .9754 | .9759 | .9763 | .9768 | .9773 |
| .5  | .9777 | .9782 | .9786 | .9791 | .9795 | .9800 | .9805 | .9809 | .9814 | .9818 |
| .6  | .9823 | .9827 | .9832 | .9836 | .9841 | .9845 | .9850 | .9854 | .9859 | .9863 |
| .7  | .9868 | .9872 | .9877 | .9881 | .9886 | .9890 | .9894 | .9899 | .9903 | .9908 |
| .8  | .9912 | .9917 | .9921 | .9926 | .9930 | .9934 | .9939 | .9943 | .9948 | .9952 |
| .9  | .9956 | .9961 | .9965 | .9969 | .9974 | .9978 | .9983 | .9987 | .9991 | .9996 |

logarithms of numbers from 100 to 1,000; Jones, G. W., giving six-place logarithms of numbers up to 10,000, in *Logarithmic Tables*; and Bauschinger and Peters, *Logarithmic Tables*, giving eight-figure logarithms of numbers up to 200,000.



## APPENDIX C

### THE METRIC SYSTEM <sup>1</sup>

The metric system of weights and measures is based on the decimal system. At the present time it is in use in most civilized countries. The system was officially adopted by the United States in 1866, but is as yet used only in scientific work. All of the metric system units of measure are based on the meter, which is the fundamental unit. It is the equivalent of 39.37 inches. Formerly, the meter was defined as one ten-millionth of the distance from the equator to the pole, but now as the distance between two lines on a certain metallic rod preserved in the archives of the International Metric Commission at Paris. In addition to the measures of length, the metric system includes also measures of surface, of which the *are* is the unit; measures of capacity, of which the *liter* is the unit; and measures of weights, of which the *gram* is the unit. The *gram* is the weight of 1 cu. cm. of distilled water in a vacuum, at its greatest density (39.2° F.).

#### LINEAR MEASURE

|                      |                       |
|----------------------|-----------------------|
| 10 millimeters (mm.) | = 1 centimeter (cm.)  |
| 10 centimeters       | = 1 decimeter (dm.)   |
| 10 decimeters        | = 1 meter (m.)        |
| 10 meters            | = 1 decameter (dcm.)  |
| 10 decameters        | = 1 hectometer (hm.)  |
| 10 hectometers       | = 1 kilometer (km.)   |
| 10 kilometers        | = 1 myriameter (mym.) |

#### SQUARE MEASURE

|                                  |                                 |
|----------------------------------|---------------------------------|
| 100 square millimeters (sq. mm.) | = 1 square centimeter (sq. cm.) |
| 100 square centimeters           | = 1 square decimeter (sq. dm.)  |
| 100 square decimeters            | = 1 square meter (sq. m.)       |
| 100 square meters                | = 1 square decameter (sq. dcm.) |
| 100 square decameters            | = 1 square hectometer (sq. hm.) |
| 100 square hectometers           | = 1 square kilometer (sq. km.)  |

<sup>1</sup> Adapted largely from *Accountants' Handbook*, The Ronald Press Company, 1923, pages 169 to 171.

## LAND MEASURE

|                    |                   |                 |
|--------------------|-------------------|-----------------|
| 100 centares (ca.) | = 1 are (a.)      | = 100 sq. m.    |
| 100 ares           | = 1 hectare (Ha.) | = 10,000 sq. m. |

## CUBIC MEASURE

|                                   |                                |
|-----------------------------------|--------------------------------|
| 1,000 cubic millimeters (cu. mm.) | = 1 cubic centimeter (cu. cm.) |
| 1,000 cubic centimeters           | = 1 cubic decimeter (cu. dm.)  |
| 1,000 cubic decimeters            | = 1 cubic meter (cu. m.)       |

## MEASURES OF CAPACITY

|                      |                      |
|----------------------|----------------------|
| 10 milliliters (ml.) | = 1 centiliter (cl.) |
| 10 centiliters       | = 1 deciliter (dl.)  |
| 10 deciliters        | = 1 liter (l.)       |
| 10 liters            | = 1 decaliter (dal.) |
| 10 decaliters        | = 1 hectoliter (hl.) |
| 10 hectoliters       | = 1 kiloliter (kl.)  |

## MEASURES OF WEIGHT

|                     |                     |
|---------------------|---------------------|
| 10 milligrams (mg.) | = 1 centigram (cg.) |
| 10 centigrams       | = 1 decigram (dg.)  |
| 10 decigrams        | = 1 gram (g.)       |
| 10 grams            | = 1 decagram (deg.) |
| 10 decagrams        | = 1 hectogram (hg.) |
| 10 hectograms       | = 1 kilogram (kg.)  |
| 10 kilograms        | = 1 myriagram (mg.) |
| 10 myriagrams       | = 1 quintal (q.)    |
| 10 quintals         | = 1 tonneau (t.)    |

## TABLES OF EQUIVALENTS

## LINEAR MEASURE

|                            |                                 |
|----------------------------|---------------------------------|
| 1 inch = 2.54 centimeters  | 1 centimeter = .3937 of an inch |
| 1 foot = .3048 of a meter  | 1 decimeter = .328 of a foot    |
| 1 yard = .9144 of a meter  | 1 meter = 1.0936 yards          |
| 1 rod = 5.029 meters       | 1 decameter = 1.9884 rods       |
| 1 mile = 1.6093 kilometers | 1 kilometer = .62137 of a mile  |

## SURFACE MEASURE

|                                    |                                       |
|------------------------------------|---------------------------------------|
| 1 sq. inch = 6.452 sq. centimeters | 1 sq. centimeter = .155 of a sq. inch |
| 1 sq. foot = .0929 of a sq. meter  | 1 sq. decimeter = .1076 of a sq. foot |
| 1 sq. yard = .8361 of a sq. meter  | 1 sq. meter = 1.196 sq. yards         |
| 1 sq. rod = 25.293 sq. meters      | 1 are = 3.954 sq. rods                |
| 1 acre = 40.47 ares                | 1 hectare = 2.471 acres               |
| 1 sq. mile = 259 hectares          | 1 sq. kilometer = .3861 of a sq. mile |

## CUBIC MEASURE

|                                     |                                       |
|-------------------------------------|---------------------------------------|
| 1 cu. inch = 16.387 cu. centimeters | 1 cu. centimeter = .061 of a cu. inch |
| 1 cu. foot = 28.317 cu. decimeters  | 1 cu. decimeter = .0353 of a cu. foot |
| 1 cu. yard = .7646 of a cu. meter   | 1 cu. meter = 1.308 cu. yards         |
| 1 cord = 3.624 steres               | 1 stere = .2759 of a cord             |

## MEASURES OF CAPACITY

|                                        |                                     |
|----------------------------------------|-------------------------------------|
| 1 dry quart = 1.101 liters             | 1 liter = .908 of a dry quart       |
| 1 liquid quart = .9463 of a liter      | 1 liter = 1.0567 liquid quarts      |
| 1 liquid gallon = .3785 of a decaliter | 1 decaliter = 2.6417 liquid gallons |
| 1 peck = .881 of a decaliter           | 1 decaliter = 1.135 pecks           |
| 1 bushel = .3524 of a hectoliter       | 1 hectoliter = 2.8377 bushels       |

## MEASURES OF WEIGHT

|                                                   |
|---------------------------------------------------|
| 1 grain, troy = .0648 of a gram                   |
| 1 ounce, troy = 31.104 grams                      |
| 1 ounce, avoird. = 28.35 grams                    |
| 1 lb., troy = .3732 of a kilogram                 |
| 1 lb., avoird. = .4536 of a kilogram              |
| 1 ton (short) = .9072 of a tonneau, or metric ton |
| 1 gram = 15.432 grains, troy                      |
| 1 gram = .03215 of an ounce, troy                 |
| 1 gram = .03527 of an ounce, avoird.              |
| 1 kilogram = 2.679 lb., troy                      |
| 1 kilogram = 2.2046 lb., avoird.                  |
| 1 tonneau = 1.1023 tons (short)                   |

## CONVENIENT EQUIVALENT VALUES

|                                                          |               |
|----------------------------------------------------------|---------------|
| 1 cu. cm. of water = 1 ml. of water, and weighs 1 gram   | = 15.432 grs. |
| 1 cu. dm. of water = 1 liter of water, and weighs 1 kg.  | = 2.2047 lbs. |
| 1 cu. m. of water = 1 kl. of water, and weighs 1 tonneau | = 2204.6 lbs. |

# APPENDIX D

TABLE XCII. SQUARES OF DEVIATIONS FROM POINTS OF ORIGIN IN STRAIGHT LINES OF LEAST SQUARES

| 1                                     | 2                                                   | 3                                                                                           | 4                                     | 5                                                   | 6                                                                                           |
|---------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------|---------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------|
| DEVIATIONS<br>FROM POINT<br>OF ORIGIN | SQUARES OF<br>DEVIATIONS<br>FROM POINT<br>OF ORIGIN | CUMULATIVE SUMMA-<br>TION OF SQUARES<br>OF DEVIATIONS,<br>BEGINNING WITH<br>POINT OF ORIGIN | DEVIATIONS<br>FROM POINT<br>OF ORIGIN | SQUARES OF<br>DEVIATIONS<br>FROM POINT<br>OF ORIGIN | CUMULATIVE SUMMA-<br>TION OF SQUARES<br>OF DEVIATIONS,<br>BEGINNING WITH<br>POINT OF ORIGIN |
| -25                                   | 625                                                 | 5525                                                                                        | -24.5                                 | 600.25                                              | 5206.25                                                                                     |
| -24                                   | 576                                                 | 4900                                                                                        | -23.5                                 | 552.25                                              | 4606.00                                                                                     |
| -23                                   | 529                                                 | 4324                                                                                        | -22.5                                 | 506.25                                              | 4053.75                                                                                     |
| -22                                   | 484                                                 | 3795                                                                                        | -21.5                                 | 462.25                                              | 3547.50                                                                                     |
| -21                                   | 441                                                 | 3311                                                                                        | -20.5                                 | 420.25                                              | 3085.25                                                                                     |
| -20                                   | 400                                                 | 2870                                                                                        | -19.5                                 | 380.25                                              | 2665.00                                                                                     |
| -19                                   | 361                                                 | 2470                                                                                        | -18.5                                 | 342.25                                              | 2284.75                                                                                     |
| -18                                   | 324                                                 | 2109                                                                                        | -17.5                                 | 306.25                                              | 1942.50                                                                                     |
| -17                                   | 289                                                 | 1785                                                                                        | -16.5                                 | 272.25                                              | 1636.25                                                                                     |
| -16                                   | 256                                                 | 1496                                                                                        | -15.5                                 | 240.25                                              | 1364.00                                                                                     |
| -15                                   | 225                                                 | 1240                                                                                        | -14.5                                 | 210.25                                              | 1123.75                                                                                     |
| -14                                   | 196                                                 | 1015                                                                                        | -13.5                                 | 182.25                                              | 913.50                                                                                      |
| -13                                   | 169                                                 | 819                                                                                         | -12.5                                 | 156.25                                              | 731.25                                                                                      |
| -12                                   | 144                                                 | 650                                                                                         | -11.5                                 | 132.25                                              | 575.00                                                                                      |
| -11                                   | 121                                                 | 506                                                                                         | -10.5                                 | 110.25                                              | 442.75                                                                                      |
| -10                                   | 100                                                 | 385                                                                                         | -9.5                                  | 90.25                                               | 332.50                                                                                      |
| -9                                    | 81                                                  | 285                                                                                         | -8.5                                  | 72.25                                               | 242.25                                                                                      |
| -8                                    | 64                                                  | 204                                                                                         | -7.5                                  | 56.25                                               | 170.00                                                                                      |
| -7                                    | 49                                                  | 140                                                                                         | -6.5                                  | 42.25                                               | 113.75                                                                                      |
| -6                                    | 36                                                  | 91                                                                                          | -5.5                                  | 30.25                                               | 71.50                                                                                       |
| -5                                    | 25                                                  | 55                                                                                          | -4.5                                  | 20.25                                               | 41.25                                                                                       |
| -4                                    | 16                                                  | 30                                                                                          | -3.5                                  | 12.25                                               | 21.00                                                                                       |
| -3                                    | 9                                                   | 14                                                                                          | -2.5                                  | 6.25                                                | 8.75                                                                                        |
| -2                                    | 4                                                   | 5                                                                                           | -1.5                                  | 2.25                                                | 2.50                                                                                        |
| -1                                    | 1                                                   | 1                                                                                           | -.5                                   | .25                                                 | .25                                                                                         |
| 0                                     | 0                                                   | 0                                                                                           | 0                                     | 0                                                   | 0                                                                                           |

TABLE XCII. SQUARES OF DEVIATIONS FROM POINTS OF ORIGIN IN STRAIGHT LINES OF LEAST SQUARES—Continued

| 1                                     | 2                                                   | 3                                                                                           | 4                                     | 5                                                   | 6                                                                                           |
|---------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------|---------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------|
| DEVIATIONS<br>FROM POINT<br>OF ORIGIN | SQUARES OF<br>DEVIATIONS<br>FROM POINT<br>OF ORIGIN | CUMULATIVE SUMMA-<br>TION OF SQUARES<br>OF DEVIATIONS,<br>BEGINNING WITH<br>POINT OF ORIGIN | DEVIATIONS<br>FROM POINT<br>OF ORIGIN | SQUARES OF<br>DEVIATIONS<br>FROM POINT<br>OF ORIGIN | CUMULATIVE SUMMA-<br>TION OF SQUARES<br>OF DEVIATIONS,<br>BEGINNING WITH<br>POINT OF ORIGIN |
| 0                                     | 0                                                   | 0                                                                                           | 0                                     | 0                                                   | 0                                                                                           |
| 1                                     | 1                                                   | 1                                                                                           | .5                                    | .25                                                 | .25                                                                                         |
| 2                                     | 4                                                   | 5                                                                                           | 1.5                                   | 2.25                                                | 2.50                                                                                        |
| 3                                     | 9                                                   | 14                                                                                          | 2.5                                   | 6.25                                                | 8.75                                                                                        |
| 4                                     | 16                                                  | 30                                                                                          | 3.5                                   | 12.25                                               | 21.00                                                                                       |
| 5                                     | 25                                                  | 55                                                                                          | 4.5                                   | 20.25                                               | 41.25                                                                                       |
| 6                                     | 36                                                  | 91                                                                                          | 5.5                                   | 30.25                                               | 71.50                                                                                       |
| 7                                     | 49                                                  | 140                                                                                         | 6.5                                   | 42.25                                               | 113.75                                                                                      |
| 8                                     | 64                                                  | 204                                                                                         | 7.5                                   | 56.25                                               | 170.00                                                                                      |
| 9                                     | 81                                                  | 285                                                                                         | 8.5                                   | 72.25                                               | 242.25                                                                                      |
| 10                                    | 100                                                 | 385                                                                                         | 9.5                                   | 90.25                                               | 332.50                                                                                      |
| 11                                    | 121                                                 | 506                                                                                         | 10.5                                  | 110.25                                              | 442.75                                                                                      |
| 12                                    | 144                                                 | 650                                                                                         | 11.5                                  | 132.25                                              | 575.00                                                                                      |
| 13                                    | 169                                                 | 819                                                                                         | 12.5                                  | 156.25                                              | 731.25                                                                                      |
| 14                                    | 196                                                 | 1015                                                                                        | 13.5                                  | 182.25                                              | 913.50                                                                                      |
| 15                                    | 225                                                 | 1240                                                                                        | 14.5                                  | 210.25                                              | 1123.75                                                                                     |
| 16                                    | 256                                                 | 1496                                                                                        | 15.5                                  | 240.25                                              | 1364.00                                                                                     |
| 17                                    | 289                                                 | 1785                                                                                        | 16.5                                  | 272.25                                              | 1636.25                                                                                     |
| 18                                    | 324                                                 | 2109                                                                                        | 17.5                                  | 306.25                                              | 1942.50                                                                                     |
| 19                                    | 361                                                 | 2470                                                                                        | 18.5                                  | 342.25                                              | 2284.75                                                                                     |
| 20                                    | 400                                                 | 2870                                                                                        | 19.5                                  | 380.25                                              | 2665.00                                                                                     |
| 21                                    | 441                                                 | 3311                                                                                        | 20.5                                  | 420.25                                              | 3085.25                                                                                     |
| 22                                    | 484                                                 | 3795                                                                                        | 21.5                                  | 462.25                                              | 3547.50                                                                                     |
| 23                                    | 529                                                 | 4324                                                                                        | 22.5                                  | 506.25                                              | 4053.75                                                                                     |
| 24                                    | 576                                                 | 4900                                                                                        | 23.5                                  | 552.25                                              | 4606.00                                                                                     |
| 25                                    | 625                                                 | 5525                                                                                        | 24.5                                  | 600.25                                              | 5206.25                                                                                     |

This table of squares and cumulative summations is designed to facilitate the calculation of ordinates of straight line trends of least squares by the method as outlined in Tables LXIII and LXIV, pages 172 and 174. Column 1 shows the deviations from the point of origin when there is an odd number of items, and column 4 shows the deviations from the point of origin when there is an even number of items. The figures in columns 2 and 5 are the squares of the deviations from points of origin, or the squares of the deviations in columns 1 and 4. The cumulative summations of these squared deviations are recorded in columns 3 and 6. To illustrate how this table may be used, let us assume a straight line of least squares is to be fitted to a series containing 25 items. Immediately it will be seen that there are 12 items on either side of the central point of origin. Referring to column 3 of the table it is seen further that the summation of the squares of these 12 deviations preceding the point of origin is 650, and that the summation of the squares of the 12 deviations below the point of origin is 650. Combining these two summations gives a total of 1300, which is the total summation of the squares of all the deviations, both plus and minus, from the point of origin. Thus, the table may be used in reducing the amount of work involved in fitting straight line trends of least squares.



## APPENDIX E

### MEASUREMENTS <sup>1</sup>

An *angle* is formed by two lines which meet.

A *right angle* is one formed by two lines which meet at an angle of 90 degrees.

A *polygon* is a plane surface bounded by straight lines.

A *quadrilateral* is a polygon bounded by four sides and having four angles.

A *parallelogram* is a quadrilateral whose opposite sides are parallel.

A *rectangle* is a parallelogram having right angles.

A *square* is a rectangle whose four sides are equal.

A *triangle* is a surface having three sides and three angles.

A *right-angled triangle* has one right angle. The side opposite the right angle is the *hypotenuse*.

An *isosceles triangle* has two sides that are equal.

An *equilateral triangle* has all sides and all angles equal.

A *circle* is a plane surface bounded by a curved line, each point of which is equidistant with every other point thereof from an interior point known as the center. The bounding line is the *circumference*. The *diameter* is a straight line passing through the center and terminating at each end in the circumference. One-half the diameter is the *radius*.

The *perimeter* of a surface is the distance around it. Its *area* is the number of square units of area which it contains.

*To find the area of a parallelogram:* Multiply the base by the altitude.

*To find the area of a triangle:* Take one-half the product of the base and altitude, or multiply the base by one-half the altitude.

<sup>1</sup> Adapted from *Accountants' Handbook*.

*To find the hypotenuse of a right triangle:* Extract the square root of the sum of the squares of base and perpendicular.

*To find the base or perpendicular of a right-angle triangle:* Extract the square root of the difference between the squares of the hypotenuse and the other given side.

*To find the circumference of a circle:* Multiply the diameter by 3.1416.

*To find the diameter of a circle:* Divide the circumference by 3.1416.

*To find the area of a circle:* Multiply the square of the diameter by .7854.

Or, multiply the circumference by one-half the radius.

Or, multiply the square of the radius by 3.1416.

*To find the diameter of a circle when the area is known:* Divide the area by .7854 and extract the square root of the quotient.

*To find the circumference of a circle when the area is known:* Multiply the diameter by 3.1416. Find diameter by preceding rule.

A solid whose bases are equal polygons lying in parallel planes and whose sides are parallelograms is called a *prism*.

A prism having six equal square faces is a *cube*.

A solid whose bases are equal and parallel surfaces and whose surface is a uniform curve is a *cylinder*.

A *pyramid* is a solid whose base is a polygon and whose sides are triangles.

A *cone* is a solid whose base is a circle and whose surface tapers to a point.

*To find the volume of a prism or cylinder:* Multiply the area of the base by the altitude.

*To find the surface of a right prism or cylinder:* Multiply the perimeter of the base by the altitude. Add to the product the area of the bases.

*To find the volume of a pyramid or cone:* Take one-third of the product of the area of the base and the altitude.

*To find the surface of a pyramid or cone:* Multiply the perimeter of the base by one-half of the slant height. Add the area of the base.

The *areas* of similar surfaces are to each other as the squares of their *dimensions*.

The *volumes* of similar solids are to each other as the cubes of their like dimensions.

*To find the area of the surface of a sphere:* The area of a sphere equals the area of four great circles of the sphere. The area of a great circle is 3.1416 times the radius squared.

*To find the volume of a sphere:* Multiply the area of its surfaces by one-third of the radius.



## INDEX





## INDEX

- Accountant's Handbook, 293n, 297n, 299n, 305n
- Accuracy of index numbers, 283-284
- Adequacy in sampling, 12-13
- Aggregates, 61
- Agricultural products, duty on, 28-29
- Agriculture, U. S. Dept. of, 5, 6, 38, 273
- Alienation, coefficient of, 276-277
- Alignment of commas and decimals, 43
- Allocation of costs, 27
- Antilogarithm, 204n
- Apples, production in the United States, 37
- Application of the probable error, 156-161
- Approximate median, 121
- Areas of the probability surface, 131n, 290-292
- Arithmetic mean, 22, 23, 95-111, 112, 117, 122, 125, 142, 143, 146, 152, 156, 157, 158, 200 (*See* Simple arithmetic mean)
  - calculation of, 97-99
  - calculation of from an assumed mean, 98-99, 108-110, 135-137
  - correction factor for, 99, 109-110, 135, 136, 137, 138, 240-242
  - definition of, 96,
  - mean of a frequency distribution, 106-110
  - nature of, 97
  - unweighted mean, 104
  - use in measuring skewness, 152
  - weighted mean, 22, 23, 24, 99-106, 107
    - calculation of, 99-106
    - meaning of, 101
- Arrangement of data, 31, 43, 53
- Array, 51, 116, 117, 118, 122, 124, 148, 149, 150
  - as a measure of variability and dispersion, 124
  - definition of, 50
- Association, degree of, 262
- Assumed mean, 98-99, 108, 109, 135, 136
  - calculation of true mean from, 98-99, 108-110, 135-137
- Asymmetrical distribution, 63-64
- Asymmetry, 64n (*See* Skewness)
- Automobile fatalities in Washington, 67-68
- Automobiles,
  - sampling of, 18-19
- Average, arithmetic (*See* Arithmetic mean)
- Average deviation, 110, 124-128, 129, 143, 150
  - as a correction factor in calculating the mean, 110
  - calculation of, 124-128
  - in the analysis of economic data, 143
  - meaning of, 125
- Average price, 6
- Average returns, highest, 89n, 91-92
- Averages, 95
- Axis of symmetry, 61
- Axis, x and y (*See* x axis and y axis)
- Baltimore,
  - estimate of population, 24
  - precipitation, 98
- Band chart, constituent-element, 76
- Bank deposits in the United States, 279

- Bar charts, 66-67
- Bauschinger and Peters, Logarithmic Tables, 298n
- Bias in sampling, 157
- Binomial curve, 139
- Binomial expansion, 64n, 139
- Biology, 95, 248
- Buckwheat, imports and exports of, 78
- Bureau of Labor Statistics, 223n, 284n
- Calculating machines, 171, 210n
- California,
  - rice acreage, 70
- Cantaloupes,
  - crates for, 6-7
  - prices, 6-7
- Capital issues of corporations, 38
- Captions, 36, 42
- Causal factors, 196
- Causal relationship, 274
- Cause and effect relationship, 195, 250, 262, 264, 275
- Center of symmetry, 61
- Central location, 123
- Central tendency, 96, 122, 124, 125, 129, 130n
- Chance error, 152
- Chance selection, 8-9, 158
- Chance variation, 25
- Characteristic of logarithms,
  - meaning of, 293
- Charts (*See* Graphic presentation)
- Circle charts, 71-72
- Citric acid, imports of, 60
- Classification of data, 32-35
  - geographical, 33-34
  - qualitative, 33
  - quantitative, 34-35
- Class frequency,
  - meaning of, 46-47
- Class interval, 48-49, 135, 136
  - magnitude of, 48-49
  - meaning of, 48-49
- Class limits,
  - meaning of, 49-50
- Class mark defined, 50
- Clothing,
  - price changes of, 6-7
- Coefficients,
  - alienation, 276-277
  - correlation (*See* Correlation)
  - determination, 228
  - dispersion, 128-129
  - regression, 249, 251, 252, 264, 275 (*See* Regression)
  - variation, 145-146, 221n
- Coincidental relationship, 195
- Coin tossing, 143
- College Park, Maryland,
  - location of, 257n
  - weather conditions, 225, 257
  - wheat yields, 225, 257
- Colleges, universities, and professional schools in the United States, 98
- Column diagram, 86
- Column headings, 40, 42
- Columns, 32, 36, 37, 42, 43, 87, 88
- Commas, 43, 44
- Commission, U. S. Tariff, 40, 58n, 72n
- Composite index number, 280-283
- Concomitant deviations,
  - coefficient of, 203-205
- Concurrent deviations,
  - coefficient of, 203-205
- Consistent relationship, 275
- Constant returns, 89n, 93-94
- Constituent-element chart, 74-76
- Construction of tables, 35-44
  - rules for, 40, 42-44
- Continuous data, 106, 114, 115
- Continuous frequency distribution, 107
- Continuous variable, 46
- Corn,
  - acreage in North Central States, 101
  - acreage on selected farms in Georgia and Indiana, 192, 193, 237
  - production in North Central States, 101

Corn—*Continued*

- relation between yield and acreage, 193, 237
- yield per acre in North Central States, 100, 101
- yield per acre on selected farms in Georgia and Indiana, 192

Corporations, capital issues of, 38

Correction factor,

- for true arithmetic mean, 99, 110, 135, 136, 137, 138, 240-242
- for true mean of a frequency distribution, 109-110
- for true standard deviation, 141-142, 144, 243-244

Correction for differences between samples, 144

Correlation, 48, 195-247, 251, 252, 255, 256, 268, 274, 275, 276

- calculation of coefficient, 203-247
- causal factors, 196
- correlation index, 233-234, 256
- correlation ratio, 234-236, 256
- correlation table, 48n, 201, 202, 237

curvilinear relationship, 233n

degree of association, 262

fundamental principle of, 196-197

index of correlation, 233-234, 256

interpretation of coefficient of correlation, 197-198, 207-208, 225, 231

inverse, 197 (*See* Inverse relationship and Negative correlation)

inverse relationship, 199, 200

meaning of coefficient, 197-198, 207-208, 225, 231

multiple correlation, 221-228

nature of, 195-197

negative, 196, 197 (*See* Inverse relationship)

of actual values, 211-212

of a frequency distribution, 238-246

of concomitant deviations, 203-205

of concurrent deviations, 203-205

Correlation—*Continued*

of deviations from trend, 212-213

of first differences, 213-216

of multiples of standard deviation, 208-211, 216-217

of percentage changes of first differences, 216-220, 252

partial correlation, 228-231

Pearsonian coefficient, 205-208

positive, 196, 197

probable error of coefficient, 160-161, 220

product-moment method, 205

rank-order correlation, 231-233

simple correlation, 202, 229n

methods, 203-220

symbol for coefficient, 197, 229n

sum-product method, 205

symbol for coefficient, 197, 225, 229n

time lag, 246-247

unreliability of coefficient in predictive equations, 274-275

Cost of living index, 282-283, 284

Costs,

allocation of, 27

Cotton,

acreage harvested, 221-224, 265-268

estimates of acreage, 265-268

estimates of production, 269-272

par forecasts, 272-274

gins, 11n

par forecasts of production, 272-274

partial correlation between supply and price, 228-231

prices, 190, 191, 198, 199, 202, 214-217, 221-224, 233

production, 190, 191, 198, 199, 202, 214-217, 233, 271, 273

estimates of, 270-271

par forecasts, 273, 274

rank-order correlation between production and price, 231-233

ratio estimates of production, 269-271

Cotton—*Continued*

- relation between price and acreage, 221–225, 264
  - relation between production and price, 190–191, 198–199, 214–215, 216–217, 231–233
  - relation between supply and price, 228–231
  - yield groups, 49
  - yield per acre,
    - in foreign countries, 39
    - in the United States, 39, 270
- Crop Reporting Board, 272n
- Cross-hatch circle chart, 71
- Cross-hatch map, 68–69
- Cumulative distribution, 54–55
- Curve,
  - binomial, 139
  - constant returns, 89n, 93–94
  - decreasing returns, 89n
  - demand, 66n
  - diminishing returns, 90
  - fixed returns, 93n
  - Gaussian, 154
  - highest average returns, 89n, 91–92
  - increasing returns, 89n, 93
  - normal-frequency, 65, 139, 151, 154–156
    - point of inflection, 131
  - of averages, 145–146
  - of distribution, 151, 155
  - of error, 154
  - of estimates, 275
  - probability, ordinates of, 287–289
  - supply, 66n
  - utility, 66n
- Curvilinear regression, 256
- Curvilinear relationship, 233n
- Cycle,
  - elimination by free-hand method, 166–167
  - relation of moving average to, 164
- Cyclical movements, 162
- Day, E. E., 111n, 284n
- Deciles, 151
- Decimal points, 43, 44

- Decreasing returns, 89n (*See* Diminishing returns)
- Deflated costs, 188
- Deflated prices, 223n, 264n
- Deflation of prices, 187–188
- Degree of association, 262
- Degree of determination, 224
- Delaware,
  - acreage of hay crops, 68–70
  - cantaloupe prices, 6–7
- Demand curve, 66n
- Dependability of index numbers, 283 (*See* Accuracy of index numbers)
- Dependent series, 196, 234, 263
- Dependent variable, 196, 221, 221n, 225, 229, 256, 262, 263, 269
- Determination, coefficient of, 228
- degree of, 224
- Deviation,
  - average, 110, 124–128, 129, 143, 150
  - mean, of a frequency distribution, 126–128
  - mean, of a series, 124–126
  - quartile, 150, 157
  - standard (*See* Standard deviation)
- Deviations,
  - squares of from points of origin in straight line trends, 302–304
- Diminishing returns, 89–91
- Discrete data, 107
- Discrete frequency distribution, 107
- Discrete series, 108, 116
- Discrete variable, 46
- Dispersion, 122–153, 234, 236
  - coefficient of, 128–129
- Distorted normal distribution, 151
- Distribution, 49, 70
  - asymmetrical, 63–64
  - cumulative, 54–55
  - distorted normal, 151
  - frequency, 35, 45–65
  - normal, 63, 64–65, 108, 117, 128, 131, 139, 145, 151, 152, 154, 155, 156, 157, 161, 290

- Distribution—*Continued*  
 order, 50  
 probable error of, 161  
 simple, 51-54  
 spatial, 55-58  
 symmetrical, 61-63, 108  
 temporal, 58-61
- Division by use of logarithms, 295-296
- Dot maps, 57-58, 70-71
- Duty on agricultural products, 28-29
- Earnings of federal reserve banks, 46
- Economics, 3, 66n, 95
- Education, 3, 5, 95
- Error (*See* Probable error)  
 chance, 152  
 curve of, 154  
 normal law of, 142-143
- Error of estimate, 275 (*See* Standard error of estimate and Coefficient of alienation)
- Estimate, 262n  
 standard error of, 274-276, 277
- Estimates,  
 curve of, 275  
 of cotton acreage, 265, 266, 268  
 of cotton production, 270-271  
 of wheat yields, 259-261
- Estimates and error, 262-277  
 unreliability of estimates, 275
- Estimating,  
 par forecasting, 272-274  
 ratio method of, 269-272
- Estimating equation, 263-269  
 reliability and unreliability of, 272, 275
- Ezekiel, M. J. B., 247n
- Excess and deficit chart, 77-78
- Factor-reversal test, 283-284
- Families, estimating size of, 9-10
- Farm cost data, 5
- Farm cost studies, 19-21
- Farm costs, 25, 26-27, 154, 155
- Farm wages, 188  
 indices of, 188
- Federal reserve bank earnings, 46
- Fertilizer, 3, 4, 10
- First degree parabola, 168, 186  
 (*See* Straight line of least squares)
- First differences,  
 correlation of, 213-216  
 multiples of standard deviation of, 86  
 percentage change method of correlation, 216-220, 252
- Fisher, Irving, 283, 284n
- Fitting of curved lines, 178-187  
 logarithmic parabola, 178-181  
 second degree parabola, 181-185  
 third degree parabola, 187  
 trends of higher order, 185-187
- Fitting of straight lines of least squares, 168-178
- Fitting of trends, 163-188
- Fixed returns, 93n
- Flax,  
 acreage in the United States, 282
- Flaxseed,  
 index of yield, 281, 282  
 production, 282  
 yield per acre, 281
- Food,  
 price changes of, 6-7
- Footnote, 36, 37, 42, 43
- Forecasting, 262n  
 par method of, 272-274 (*See* Estimates and Estimating)
- Free-hand trend, 163-164, 166
- Frequency, class, 47
- Frequency column, 53
- Frequency curve (*See* Normal frequency curve)
- Frequency distribution, 35, 45-65,  
 105, 112, 113, 123, 124, 126,  
 132, 133, 134, 149, 202, 239  
 asymmetrical, 63-64  
 continuous, 107  
 correlation of, 238-246  
 cumulative, 54-55  
 definition of, 48  
 definition of terms, 45-51  
 discrete, 107  
 mean deviation of, 126-128



Frequency—*Continued*

mean of, 106-110, 127

calculated from an assumed  
mean, 108-110

meaning of, 48

normal, 64-65 (*See* Normal dis-  
tribution)

simple, 51-54

spatial, 55-58

standard deviation calculated  
from, 132-143

symmetrical, 61-63

temporal, 58-61

## Frequency graph, 86-88

as an indication of variability and  
dispersion, 123-124

## Frequency polygon, 65, 87-88

## Frequency surface,

meaning of, 48

Frequency table, 48, 53, 54, 123,  
238, 246as an indication of variability  
and dispersion, 123

nature of, 48

## Gasoline, mileage tests for, 5

## Gauss, Karl Friedrich, 154

## Gaussian curve, 154

## Geneticist, 11-12

## Geographical classification, 33-34

## Geometric mean, 110-111

## Georgia,

acreage and yield of corn, 192,  
193, 237

## Graphic presentation, 56, 60, 66-94

band chart, 76

bar charts, 66-67

circle charts, 71-72

column diagram, 86

constant returns, 93-94

constituent-element chart, 74-76

diminishing returns, 89-91

excess and deficit chart, 77-78

fixed returns, 93n

frequency graphs, 86-88

histogram, 86-87

polygon, 87-88

smoothed curve, 88

Graphic presentation—*Continued*

highest average returns, 91-92

histogram, 86-87

historigram, 72-74, 80, 86

increasing returns, 92-93

maps, 68-71

meaning of, 66

polygon, 87-88

ratio chart, 79-81

regression chart, 84-86

scatter diagram, 82-83

silhouette chart, 75-77

smoothed curve, 88

## Graphic representation, 48

## Grouped data, 51

## Grouping of items, 45

Handbook of Mathematical Statis-  
tics, 161n

## Harmonic mean 110-111

## Hay crops in Delaware, 68-69

Hay, sorghum, illustrative data on  
costs, 105, 109

## Headings, 32, 36, 37, 40, 42

## High school grades,

sampling of, 5-6

Highest average relationship, 85,  
250Highest average returns, 89n, 91-  
92

## Histogram, 86, 87, 88, 123-124

Historical series (*See* Time series)

## Historigram, 72-74, 80, 86

## Hopkins, C. G., 39n

## Horizontal bar chart, 66-67

Immigration into the United States,  
61

## Increasing returns, 92-93

## Indentation, 42-43

Independent series, 196 (*See* Inde-  
pendent variable)Independent variable, 196, 221,  
221n, 225, 229, 256, 262, 263,  
269indication of which affects the  
dependent to the greater ex-  
tent, 221n



- Index numbers, 278-284  
   accuracy of, 283-284  
   composite index, 280-283, 281n  
   definition of, 278  
   factor-reversal test, 283-284  
   link index number, 280  
   relatives, 281n  
   simple index numbers, 279-280  
   time-reversal test, 283-284  
   unweighted index number, 280-283  
   weighted index number, 280-283  
 Index of correlation, 213n, 233-234, 256  
 Indiana,  
   acreage and yield of corn, 192, 193, 237  
 Indices,  
   of bank deposits in the United States, 279  
   of cost of living, 282-283, 284  
   of farm wages, 188  
   of wheat yields in Maryland, 186-187  
   of yield of flaxseed, 281, 282  
 Industries in the United States, manufacturing, 34  
 Inertia of large numbers, 27, 28  
 International Metric Commission, 299  
 Interpolation,  
   in determining numbers corresponding to logarithms, 294  
 Interrelationship, 259n  
 Interval selection, 9-10  
 Inverse relationship, 199, 200 (*See* Negative correlation)  
  
 Jerome, Harry, 187n, 284n  
 Jones, G. W., 298n  
 Juvenile delinquency, 7-8  
  
 Kelley, T. L., 232n  
 Kemp, W. B., 247n  
 King, W. I., 114n, 119n, 197n, 203n  
 Kurtosis, 151n  
  
 Labor Statistics, U. S. Bureau of, 223n, 284n  
 Labor, U. S. Department of, 7  
 Lag, determination of, 74n  
   time lag in correlation, 246-247  
 Large numbers, inertia of, 27, 28  
 Laws (*See* Statistical laws)  
 Least squares, line of, 168-178, 255  
   (*See* Straight line of least squares)  
 Linear relationship, 189-193, 250, 254  
 Line charts, 86  
 Line of best fit, 255  
 Line of least squares, 168-178, 185, 255  
   formula for straight line, 172, 173, 176  
   limitations of, 177  
 Link index numbers, 280  
 Link relative, 280  
 Logarithmic paper, 81n  
 Logarithmic parabola, 178-181  
   second degree, 178n  
 Logarithmic scale,  
   principle of, 81n  
 Logarithmic trend, 181  
 Logarithms, 81n, 111, 204n, 293-298  
   antilogarithm, 204n  
   characteristic, 293, 294  
   definition of, 293  
   mantissa, 293-294  
   numbers corresponding to, 294  
   systems, 293  
   tables of, 297-298  
   use in division, 293, 295-296  
   use in extracting roots of numbers, 204n, 293, 295  
   use in multiplication, 206n, 293, 294-295  
 Long-time tendency, 155  
 Long-time trend, 162-188 (*See* Trend)  
   meaning of, 162  
  
 Mantissa,  
   meaning of, 293  
 Manufacturing industries in the United States, 34

- Maps, 57-58, 66, 68-71
  - cross-hatch, 68-69
  - multiple dot, 57-58, 70-71
  - single dot, 57-58, 71
- Maryland Agricultural Experiment Station, 38, 39n
  - location of, 257n
- Maryland, Eastern Shore of, 6
- Mathematical Statistics, Handbook of, 161n
- Mean (*See* Arithmetic mean)
- Mean deviation (*See* Average deviation)
  - as an indication of variability and dispersion, 124-126
  - of a frequency distribution, 126-128
  - of a series, 124-128
- Mean difference, 125
- Mean of a frequency distribution, 106-110, 127
- Measurements, 305-307
- Median, 117-121, 122, 126, 128, 146, 149, 150, 153, 156
  - approximate median, 121
  - definition of, 117
  - formula for locating, 119, 120
  - location of, 118-121
- Methods of sampling, 8-10
- Metric system, 299-301
- Midpoints of classes, 106, 107, 126n, 127
  - calculation of mean from, 105, 106-110
  - calculation of standard deviation from, 134-143
- Mills, F. C., 141n, 187n, 244n
- Mitchell's index numbers, 284n
- Modal class, 114
- Modal value, 63, 116, 155
- Mode, 63, 112-116, 117, 122, 126, 128, 151, 152, 155
  - approximation of, 114
  - definition of, 112
  - estimated mode, 115
  - formula for locating, 114
  - location of, 112, 114-116
  - use in measuring skewness, 152
- Moment, 151n
- Moving average, 164-166
- Multiple correlation, 221-228
  - meaning of, 221
  - symbol for, 225
- Multiple dot map, 57-58, 70
- Multiple regression, 256-261, 263
  - equation for, 258, 259
  - use in forecasting, 272
- Multiples of standard deviation, 86, 199, 200, 201, 208-211
  - correlation of, 208-211, 216-217
  - plotting of, 86, 200
- Multiplication by use of logarithms, 206n, 294-295
- National Fertilizer Association, 284n
- National Research Council, 161n
- Negative correlation, 196, 197 (*See* Inverse relationship)
- Negative linear relationship, 189, 191
- Net relationship, 229
- Nonlinear relationship, 193-194, 233n
- Normal distribution, 63, 64-65, 108, 117, 128, 131, 139, 145, 146, 151, 152, 154, 155, 156, 157, 161, 290
  - curve of, 65, 139
- Normal-frequency curve, 65, 139, 151, 154-156
  - nature of, 154
  - point of inflection, 131
- Normal-frequency distribution (*See* Normal distribution)
- Normal law of error, 142-143
- North Carolina, 201
- Office tabulation, 35-36
- Offsetting tendencies, 27-28
- Oil-cloth imports, 76, 77
- Order distribution, 50
- Ordinates of the probability curve, 287-289
- Overestimates, 27, 28

- Parabola,  
   first degree, 168, 186  
   logarithmic, 178-181  
   second degree, 181-185, 186  
   second degree logarithmic, 178n  
   third degree, 186-187
- Par forecasting, 272-274
- Partial correlation, 228-231  
   nature of, 228-229
- Path coefficient, 223, 224, 226, 227n,  
   228, 259
- Peanuts,  
   costs of growing, 72  
   imports of, 58-59  
   production in the United States,  
     95  
   relation between precipitation  
     and yield, 83, 85  
   yield per acre in Virginia, 84
- Pearson, Karl, 144n, 288n, 291n
- Pearsonian coefficient, 205-208, 251,  
   252  
   calculation of, 205-207  
   formula for, 205, 207  
   weaknesses or limitations of,  
     208
- Pearsonian method of correlation,  
   205-208, 255 (*See* Pearsonian  
   coefficient)
- Percentage change of first difference  
   method of correlation, 216-  
   220, 252
- Percentiles, 151
- Permanence of small numbers, 26
- Perpendiculars, 48
- Plane of symmetry, 61
- Point of inflection on a normal  
   curve, 131
- Polygon, frequency, 65, 87-88
- Population,  
   in the United States, 34, 37  
   statistical meaning of term, 11n
- Position of tables in text, 44
- Positive correlation, 196, 197 (*See*  
   Correlation)
- Positive linear relationship, 189
- Potatoes,  
   prices of, 72-73, 200, 206, 209, 239
- Potatoes—*Continued*  
   production of, 200, 206, 209, 239  
   relation between production and  
     price, 200, 206-207, 209-210,  
     239-245
- Precipitation,  
   at Baltimore, 98  
   at Norfolk, 82  
   at Portland, Ore., 98
- Precision in sampling, 29-30
- Prediction, 262n (*See* Estimates and  
   Estimating)  
   from regression equation, 263
- Predictive equation,  
   reliability of, 272  
   unreliability of, 275
- Price deflation, 187-188
- Price levels (*See* Index numbers)  
   determination of changes in, 7
- Prices of food and clothing,  
   changes in, 6-7
- Probability, 154, 250, 290
- Probability and error, 25, 154-  
   161
- Probability curve, 154  
   ordinates of, 287-289
- Probability surface,  
   areas of, 131n, 290-292
- Probable error, 30, 152-153, 158,  
   159, 160, 161, 197, 198  
   application of, 156, 161  
   as a measure of reliability, 153,  
     158, 159  
   nature of, 152  
   of a distribution, 161  
   of a single variate, 159  
   of coefficient of correlation, 160-  
     161, 220  
   of mean of a sample, 156-159  
   of standard deviation, 159-160  
   test of reliability, 158-159
- Product-moment, 211-212
- Product-moment method of correla-  
   tion, 205
- Professional schools, colleges, and  
   universities in the United  
   States, 98
- Progressive mean, 110

- Proportionate sample, 13, 17, 19, 20, 22-23, 24, 103 (*See* Weighting in sampling)
- Purchasing power of wages, 188
- Quadratic mean, 130n
- Qualitative classification, 33
- Quantitative classification, 34-35
- Quartile, 146-149, 151, 154, 159, 161  
     definition of, 146  
     formula for location, 146  
     location of in a frequency distribution, 149  
     location of in an array, 147-148
- Quartile deviation, 150, 157  
     definition of, 150
- Quartile distance, 159, 161
- Railway wheels, imports of, 41
- Random sample, 155 (*See* Random selection)
- Random selection, 8-9, 153, 156, 157, 158-159, 160, 161, 220
- Range, 51, 124, 127, 156, 157, 160, 161  
     as a measure of dispersion, 124  
     definition of, 51  
     semi-interquartile, 150
- Rank-order correlation, 231-233
- Ratio change, 81
- Ratio chart, 79-81
- Ratio method of estimating, 269-272
- Ratio paper, 79
- Reciprocal, 111  
     use in obtaining quotients, 210n
- Rectilinear regression, 251, 256
- Regression, 248-261, 262, 263, 264  
     chart, 84-86, 191, 252  
     coefficient of, 249, 251, 252, 264, 265, 275  
     curvilinear, 256  
     derivation of term, 248  
     equation, 249, 255, 258, 259, 261, 263  
     prediction from, 263  
     line of, 85, 191, 248-249, 250, 251, 256, 262, 263
- Regression—*Continued*  
     meaning of, 248  
     meaning of coefficient, 249  
     multiple, 256-261, 263  
     rectilinear, 251, 256  
     simple, 250-256  
     x on y, 252-255  
     y on x, 252-255
- Regression chart, 84-86, 191, 252
- Regression equation, 249, 255, 258, 259, 261, 263
- Regression of x on y, 252-255
- Regression of y on x, 252-255
- Regularity, 25  
     sampling as an attribute of, 24
- Relationship,  
     causal, 274  
     cause and effect, 195, 250, 262, 264, 275  
     coincidental, 195  
     consistent, 275  
     curvilinear, 233n  
     highest average, 85, 250  
     inverse 199, 200 (*See* Negative correlation)  
     linear, 189-193, 250, 254  
     measurement of (*See* Correlation)  
     negative linear, 189, 191  
     net, 229  
     nonlinear, 193-194, 233n  
     positive linear, 189
- Relative changes, 278, 279, 284  
     (*See* Index numbers)
- Relative magnitudes, 278
- Relative numbers, 279 (*See* Index numbers)
- Relatives, 281n
- Reliability, 152-153, 155, 158 (*See* Probable error)  
     of a predictive equation, 272  
     of a sample, 4, 13, 29-30, 156-159
- Representativeness, 10-13, 22-23, 103, 115
- Representative sample, 22-23, 103, 104, 131, 142  
     of economic data, 155
- Response groups, 272

- Returns,
  - constant, 89n, 93-94
  - decreasing, 89n
  - diminishing, 89-91
  - fixed, 93n
  - highest average, 89n, 91-92
  - increasing, 89n, 92-93
  - total, 92
- Reversion to type, 248
- Rice mills, illustrative data, 33
- Root-mean-square deviation, 130, 276, 277
- Roots of numbers,
  - extraction by use of logarithms, 204n, 293, 295
- Rows, 42, 43, 44
- Rules for tabulation, 40-44
- Running average, 164 (*See* Moving average)
- Rutabagas, illustrative data on costs, 104
- Saliers, E. A., 297n
- Sample (*See* Sampling)
  - adequacy of, 13
  - proportionate selection of, 13-24
  - reliability of, 13
  - representativeness in sampling, 10-13
  - size of, 12, 29-30
  - weighting of, 13-24
- Samples, correction for differences between, 144
  - correction for smallness of, 144
- Sampling, 1-30, 103
  - adequacy in, 12-13
  - application of, 5-8
  - as an attribute of regularity, 24
  - chance selection, 8
  - interval selection, 9
  - method of, 8-10
  - precision in, 29-30
  - proportionate selection in, 13-24
  - random selection, 8
  - reliability of a sample, 13
  - representativeness, 10-13
  - size of sample, 12, 29-30
- Sampling—*Continued*
  - spatial selection, 9-10
  - special selection, 10
  - statistical laws, 24-30
  - uses of, 3-8
  - weighting of a sample, 13-24
- Scatter diagram, 82-83, 254
- Seasonal variations, 187n
- Second degree logarithmic parabola, 178n
- Second degree parabola, 181-185, 186
- Second quartile, 146n
- Semi-average, 167
  - use in trend fitting, 167-168
- Semi-interquartile range, 150
- Semi-logarithmic paper, 81n
- Series, 47, 50, 162, 166, 220
  - definition of, 47
  - dependent, 196, 234, 263 (*See* Dependent variable)
  - fitting of a line to, 162
  - independent, 196 (*See* Independent variable)
  - mean deviation of, 124-126
  - time, 162, 220
- Sheppard, W. F., 144n, 288n, 291n
- Sheppard's formula, 141, 141n
- Silhouette chart, 75-77
- Simple arithmetic mean, 22, 23, 100, 101
  - calculation of, 97-99, 102, 103, 105, 111 (*See* Arithmetic mean)
- Simple correlation, 202, 229n
  - methods of, 203-220
  - symbol for coefficient, 197, 229n
- Simple distribution, 51-54
- Simple index number, 279-280
- Simple regression, 250-256
  - equation of, 255
- Single dot map, 57-58, 71
- Size of families, estimate of, 9-10
- Size of sample, 12, 29-30
- Skewed distribution, 128, 155
- Skewness, 64n, 151-152
- Small numbers, permanence of, 26
- Smith, B. B., 269n



- Smith-Hughes Act, 96n  
 Smoothed frequency curve, 88  
 Sociology, 3  
 Soil-type studies, sampling for, 9  
 Sorghum hay, illustrative data on costs, 105, 109  
 Spacings in tabulation, 36, 37, 43, 44  
 Spatial distribution, 55-58  
 Spatial selection in sampling, 9-10  
 Spearman's rank-order formula, 232, 232n  
 Special selection of data, 10, 194n, 220  
 Square root,  
   extraction by use of logarithms, 204n, 293, 295  
 Squares of deviations from points of origin in straight line trends, 302-304  
 Standard deviation, 130-145, 152, 154, 156, 157, 158, 161, 221n, 242-244, 251, 275, 276-277, 287, 290  
   as a measure of variability and dispersion, 131-132  
   as an exponent of normal-frequency, 155-156  
   as corrected for the universe, 144  
   calculation,  
     from a frequency distribution, 132-143  
     from an assumed mean, 134-138  
     from the median, 145  
     in class-interval units, 134-138  
     of a truly representative sample, 142  
     of an array or series, 130-132  
   conversion from class-interval units to actual values, 136  
   correction for differences between standard deviations of two samples, 144  
   correction for smallness of sample, 144  
   correction for when calculated from midpoints of classes, 141-142  
 Standard deviation—*Continued*  
   definition of, 130-131  
   formula for, 130, 134, 135, 135n, 138, 141, 142, 143  
   multiples of, 86, 199, 200, 201, 208-211  
     correlation of, 208-211, 216-217  
     plotting of, 86, 200  
     probable error of, 159-161  
     use in measuring skewness, 152  
 Standard error of estimate, 274-276, 277  
   definition of, 275  
 Statistical laws, 24-30  
 Statistical tables (*See* Tabulation)  
 Straight line of least squares, 168-178, 185  
   formula for, 172, 173, 176  
   limitation of, 177  
   regression chart, 85  
   regression line, 191  
   squares of deviations from origin, 302-304  
 Straight line relationship, 191, 251, 252, 255, 262, 263  
   regression chart, 85  
 Stratification, 10n  
 Subclasses, 32  
 Subheadings, 36, 42  
 Sum-product method of correlation, 205  
 Supply curve, 66n  
 Surface, areas of probability, 290-292  
 Symbol for correlation coefficient, 197, 225, 229n  
 Symmetrical distribution, 61-63, 108  
 Symmetry,  
   axis of, 61  
   center of, 61  
   plane of, 61  
 Tables,  
   captions, 36, 42  
   classification of data, 32-35  
   columns, 36, 37, 42, 43  
   of a histogram, 87, 88



- Tables—*Continued*  
 commas, 43, 44  
 construction of, 35-44  
 dash, 43  
 decimal points, 43, 44  
 footnote, 36, 37, 42, 43  
 for office use, 35-36  
 headings, 32, 36, 37, 40, 42  
 indentation, 42-43  
 number, position of, 40  
 position in text, 44  
 rows, 42, 43, 44  
 rules for construction, 40, 42-44  
 spacing, 36, 37, 43, 44  
 subclasses, 32  
 subheadings, 36, 42  
 titles, 36, 37, 44  
   position of, 40  
 Tabulation, 31-44 (*See* Tables)  
   meaning of, 31  
 Tariff Commission, U. S., 40  
 Tariff duties, 28-29  
 Taxpayers, sampling of, 10  
 Telephone subscribers, sampling of, 10  
 Temporal distribution, 58-61  
 Test for index numbers, 283-284  
 Third degree parabola, 186-187  
 Time element, 73, 86, 155  
 Time lag in correlation, 246-247  
   determination of, 74n  
 Time-reversal test, 283-284  
 Time series, 162  
   probable error of the coefficient  
     of correlation in relation to, 220  
 Titles of tables, 36, 37, 44  
   position of, 40  
 Tobacco,  
   yield per acre in the United States, 117, 163, 164, 165, 166, 172, 173, 175  
 Tolley, H. R., 247n  
 Tomatoes,  
   acreage harvested in Maryland, 148  
   acreage harvested in the United States, 74-76  
 Trends, 162-188  
   free-hand, 163-164, 166  
   logarithmic parabola, 178-181  
     second degree, 178n  
   long-time, 162  
   moving average, 164-167  
   price deflation, 187-188  
   second degree parabola, 181-185, 186  
   semi-average, 167-168  
   significance of trend, 162  
   straight line of least squares, 168-178, 185  
     formula for, 172, 173, 176  
     limitation of, 177  
   trends of higher orders, 185-187  
 Typical prices, 7  
 Underestimates, 27, 28  
 Ungrouped data, 51  
 U. S. Bureau of Labor Statistics, 223n, 284n  
 U. S. Department of Agriculture, 5, 6, 38, 273  
 Universe, 11, 18, 27, 115, 142, 152, 155, 157, 159, 160, 194n  
   statistical meaning of, 11n  
 Universities, colleges, and professional schools in the United States, 98  
 Unreliability, 152n  
   of coefficient of correlation in predictive equations, 274-275  
   of economic data, 155  
   of farm costs, 155  
   of predictive equations, 275  
 Unrepresentativeness,  
   of a sample, 103, 157  
   of economic data, 155  
   of farm costs, 155  
 Unsmoothed normal-frequency curve, 139, 140  
 Unweighted index number, 280-283  
 Unweighted mean, 104  
 Uses of sampling, 3-8  
 Utility curve, 66n

- Variable,  
   continuous, 46  
   definition of, 45  
   dependent, 196, 221, 225, 229,  
     256, 262, 263, 269  
   discrete, 46  
   independent, 196, 221, 225, 229,  
     256, 263, 269  
   meaning of, 45
- Variability, 123, 125
- Variability and dispersion, 122-153  
   coefficient of dispersion, 128-130  
   coefficient of variation, 145-146  
   deciles, 151  
   frequency graph, 123-124  
   frequency table, 123  
   mean deviation of a frequency  
     distribution, 126-128  
   mean deviation of a series, 124-  
     126  
   percentiles, 151  
   probable error, 152-153  
   quartile, 146-149  
   quartile deviation, 150  
   range, 124  
   skewness, 151-152  
   standard deviation, 130-145
- Variate frequency defined, 47
- Variate, probable error of, 159
- Variate value, 46, 48  
   definition of, 46
- Variation, 26, 99, 122, 158  
   chance, 25  
   coefficient of, 145-146, 221n  
   seasonal, 187n
- Vertical bar chart, 66-67
- Virginia,  
   precipitation at Norfolk, 82  
   scatter diagram of precipitation  
     and yield of peanuts, 83  
   yield of peanuts per acre, 84
- Vocational schools, teachers in, 96
- Wages,  
   indices of, 188  
   purchasing power of, 188
- Washington,  
   estimate of population, 24
- Weighted arithmetic mean, 22, 23-  
   24, 99-106, 107  
   calculation of, 99-106  
   meaning of, 101
- Weighted average (*See* Weighted  
   arithmetic mean)
- Weighted index number, 280-283
- Weighting, 10n  
   in sampling, 13-24
- Weights and measures, metric sys-  
   tem of, 299-301
- Wheat, 19-20  
   estimates of yield, 259-261  
   indices of yield in Maryland,  
     186-187  
   relation between weather condi-  
     tions and yield, 225-228  
   yield per acre in the United  
     States, 174, 225, 257
- x axis, 73, 81n, 82, 84, 86, 87, 92,  
   141, 167, 168, 251, 254, 287,  
   290  
   meaning of, 73
- x value, 82, 86, 87, 141, 167, 196,  
   287
- x variable, 252
- y axis, 73, 81n, 84, 86, 87, 167, 254,  
   287  
   meaning of, 73
- y ordinate, 167, 168, 287
- Yule, G. U., 141n
- y value, 82, 86, 196
- y variable, 252

















# DATE DUE

AP 16 1973

FEB 10 1972

311

36833

H23e

ELMIRA COLLEGE GANNETT-TRIPP LIBRARY

HA29 .H25 main  
Harper, Floyd Henry/Elements of practica



3 5266 00031 8341

GANNETT-TRIPP LEARNING CENTER  
ELMIRA COLLEGE

